On Counterfactual Analysis of Differentiable Functionals

YAROSLAV MUKHIN*

Abstract. Counterfactual probability distributions are important elements of policy analysis, decomposition analysis, robustness and sensitivity analysis in empirical economics. In this paper we solve two complementary problems of statistical counterfactual analysis: (i) Given a counterfactual change in a scalar functional of a probability distribution, we describe the counterfactual distributions that have such an effect on the functional and deviate minimally from the status quo distribution in a continuous fashion. (ii) Given a counterfactual distribution, we compute the change in a statistical functional relative to the status quo distribution by integrating its local changes along a path from the status quo to the counterfactual distributions. In combination, these two exercises provide a general framework for measuring the local and global relationships between (structural) estimators of parameters or counterfactuals and descriptive statistics or specific features of the data. To solve these problems, we use von Mises calculus (i.e. influence functions), information geometry, optimal transport and introduce gradient score flows. Specifically, we define a unique path of counterfactual distributions with a combination of a statistical functional and a metric of distance or cost on the nonparametric manifold of probability distributions via the gradient flow of the functional. We describe the gradient flow paths obtained with the Fisher-Rao information metric, 2-Wasserstein optimal transport metric, and their weighted variants.

*ymukhin@mit.edu

Date: October 25, 2019

I thank Victor V. Chernozhukov, Wilfrid Gangbo, Chris A.J. Klaassen (discussant), Ming Li (discussant), Anna Mikusheva, and Whitney K. Newey for their comments, support and encouragement on this project.

0. Introduction

Counterfactual distributions and their scalar functionals are important elements of policy analysis (Stock, 1989 [68], Heckman and Vytlacil, 2007 [42]) and decomposition analysis (Oaxaca, 1973 [58], Blinder, 1973 [14], Fortin, Lemieux, and Firpo, 2011 [33, 31], Chernozhukov, Fernández-Val, and Melly, 2013 [20]) in empirical economics. Canonical applications include assessing wage discrimination [58, 14, 20], predicting the effect of cleaning up a hazardous waste cite on the distribution of local house values [68, 69], estimating the effect of de-unionization on the distribution of wages [25, 30], and many others.

The idea here is to think of policy as a change in the distribution of policy variables and covariates X. The exercise is to construct a counterfactual distribution of an outcome variable Y resulting from a policy. This is done by providing (an estimate of) the counterfactual distribution of X, and assuming that the conditional distributions of Y given X remain unchanged and can be estimated from data.

The effect of policy on the distribution of Y is often quantified via a scalar functional ψ of the marginal distribution of Y, i.e. a statistic. Different functionals are used in different applications. For example, the researcher might be interested in the effect on the mean of wages [14, 58] or house prices [69], or the effects on the quantiles, variance or Gini coefficient of the logarithm of wages [33]. The goal of policy analysis is to evaluate the effect of policy on ψ. The goal of decomposition analysis is to explain the effect on ψ by attributing it to changes in distinct factors (i.e. scalar parameters) of the distribution of X. For example, the effect on ψ can be attributed to changes the means [14, 58] or location parameters [25, 30] of each component of X.

[14, 58, 68] use regression to estimate and decompose the policy effect on the mean of Y, [25] derive a local approximation to the effect on a general ψ of the policy that shifts the distribution of X, [20] estimate the entire counterfactual distribution of Y. To our knowledge, there is no systematic and computationally efficient way of evaluating the exact policy effect on a general functional ψ available in the literature. Also, there is no systematic way of constructing counterfactual distributions for changes in general functionals of X in Oaxaca-Blinder type decompositions.

Similar statistical exercises have recently been discussed for functionals ψ that are described as large sample limits of structural estimators and answer policy or counterfactual questions under the identifying assumption of the structural model. The idea here is to establish transparency of structural estimators by relating them to intuitive features of the data, and to inform the reader about the implications of the potential misspecification, mismeasurement, and violation of the identifying assumptions.

Gentzkow and Shapiro (2015) [36] report that empirical papers relying on structural models often discuss heuristically how their estimators $\widehat{\psi}$ depend on specific features of the data in order to elucidate the inner-workings of their estimators and to lend credibility to their findings. [36] propose formal *sensitivity* measures to supplement such discussions. In subsequent work, Andrews, Gentzkow and Shapiro (AGS 2017, 2018) measure the local effect on ψ of a change in the estimation moments of the distribution of the data X [6], and the sensitivity of ψ with respect to the variations in general descriptive statistics of the distribution of X [7].

In econometric terminology, the goal here is to solve the following:

Problem 0 Find the *effect* of a change in an arbitrary scalar functional ν of the distribution of the data on the value of the structural functional ψ.

AGS [36, 6, 7] and the related econometric papers, Bonhomme and Weidner (2018) [15], Armstrong and Kolesár (2018) [8], consider local (infinitesimal) changes in the distribution of

X, interpret them as misspecification and refer to the effects on ψ as "sensitivity". However, the nature of these exercises is mathematically similar to that of policy analysis [68, 69, 42] and decomposition analysis [14, 58, 25, 30, 20], because the main point is to evaluate the effect of a variation in the population distribution on a statistical functional. The novelty in [36, 6, 7] is to implicitly map a perturbation in a scalar feature ν of the distribution, to a perturbation in the entire distribution. To our knowledge, there is no systematic and computationally tractable way of specifying a local variation in the distribution with a local change in a scalar functional that has been discussed in the economics literature.

Furthermore, as noted in Christensen and Connault (2019) [21], local sensitivity analysis may fail due to nonlinearities in either ψ or ν, or a mismatch in the magnitudes of the sampling variability in $\widehat{\psi}, \widehat{\nu}$ and the misspecification. The effect of a perturbation in ν on ψ may vary with the distribution of the data. It is therefore important to solve Problem 0 exactly, that is, non-locally.

The main goal of the present paper is to propose a general econometric framework for distributional counterfactual/sensitivity analysis, that relates these literatures by solving two complementary problems:

Problem 1 Given a scalar *policy* functional ν of the probability distribution P_0 and a finite increment h, we would like to find counterfactual probability distributions $P_h = P_{\nu,h,P_0}$ that have the given effect

$$\nu(P_h) = \nu(P_0) + h, \qquad h \in J \subset \mathbb{R} \tag{0.1}$$

on the parameter ν, and deviate minimally from P_0 in a suitable sense.

The idea is to think of a counterfactual scenario (e.g. policy, decomposition, misspecification, mismeasurement, violation of identifying assumptions) as transforming a scalar statistic ν from the status quo value $\nu(P_0)$ to the counterfactual value $\nu(P_0) + h$. In order to predict the counterfactual effect on any other parameter, we need to characterize the entire counterfactual distribution P_h.

Problem 2 Given an *outcome* functional ψ of the probability distribution P_0 and a counterfactual distribution P_h, we would like to evaluate the change in ψ at P_h relative to the status quo distribution P_0.

Furthermore, the solution of Problem 2 should be applicable to the setting of Problem 1, where P_h is defined implicitly and depends on P_0. The solution of Problem 2 should also be applicable to the setting of traditional Oaxaca-Blinder decompositions, where it is assumed that the conditional distributions of Y given X do not change and the counterfactual distribution of X is specified without reference to P_0.

With the counterfactual distributions P_h specified implicitly via a change in a functional ν by solving Problem 1, we obtain the counterfactual effects:

$$\Delta_h \psi(P_0) := \psi(P_h) - \psi(P_0), \qquad h \in J \subset \mathbb{R} \tag{0.2}$$

of changes in ν on ψ, which can be evaluated by solving Problem 2. Our methodology thus solves Problem 0 and contributes to the sensitivity analysis of structural estimators by describing the global relationship between ψ and ν and by making the counterfactual distributions for finite (non-local) changes explicit. Our methodology contributes to the policy analysis by evaluating the change in an arbitrary statistic ψ directly, and to the decomposition analysis by constructing counterfactual distributions for general factors ν of the covariates X.

An important limitation in the scope of the paper is to avoid entirely the questions of

causality and endogeneity and to focus solely on descriptive methodology[2].

Methodology The policy parameter ν is a 1-dimensional object (i.e. a scalar), whereas the state of nature P_0, of which the parameter ν is a functional, is, in general, an infinite-dimensional object (i.e. a probability measure). Therefore, by a simple degree counting argument, the mapping from the counterfactual increment h in ν to the counterfactual distribution P_h is, in general, underdetermined and not one-to-one. In other words, eqs. (0.1) and (0.2) identify a *set* of counterfactual distributions and effects. This means that the researcher [e.g. 36, 6, 7] must inevitably impose infinitely more structure on Problem 1 in order to discipline partial identification and obtain a one-to-one relationship between parameters ν and ψ in eq. (0.2). This phenomenon is known as *path dependence* in the decomposition literature [e.g. 33]. Despite its popularity with the practitioners, the implicit additional structure in the sensitivity measures of [36, 6, 7] has not yet been discussed in the sensitivity analysis literature.

This paper formalizes the required additional structure in Problem 1 in a general way via a suitable infinite-dimensional parameter by using simple geometric insights. This paper also formalizes the connection between infinitesimal perturbations (and their sensitivities) and non-infinitesimal or global counterfactual distributions (and their effects) using *gradient flows* on the space of probability measures. To this end, we adopt the geometric framework of a *manifold* $\mathcal{P}$ in order to apply traditional calculus techniques (i.e. differentiation and integration) to probability distributions. Set $\mathcal{P}$ is a nonparametric collection of probability distributions on the sample space $\mathcal{X}$, which needs to be restricted only by regularity conditions that depend on the functionals ν and ψ.

The missing parameter that we introduce is a *metric* of distance $\mathfrak{g}$ on the set of counterfactual distributions $\mathcal{P}$. The purpose of the metric is to determine the *direction of steepest increase* in the policy functional

$$\nu : \mathcal{P} \to \mathbb{R}$$

at each point $P \in \mathcal{P}$. We call this direction the *gradient score* of ν at P. Because the gradient is typically unique and depends on the notion of distance on $\mathcal{P}$, the metric $\mathfrak{g}$ is a natural parameter to regularize Problem 1. Furthermore, we show how, under regularity conditions, local misspecification deviations and sensitivities can be connected and integrated to obtain global (non-infinitesimal) counterfactual distributions (0.1) and effects (0.2): We define a path of counterfactual distributions $\{P_{\mathfrak{g},h}\}_{h\in J}$ to be the trajectory of the *gradient flow* of the functional ν on $\mathcal{P}$ that passes through the status quo distribution P_0. We define the corresponding counterfactual effects $\Delta_{\mathfrak{g},h}\psi(P_0)$ by the eq. (0.2).

Our use of geometric techniques for counterfactual or policy analysis is not only technically important and convenient, but is also economically and statistically natural. The metric has an economic interpretation of the *cost* of changing the state of nature $P \in \mathcal{P}$. The direction of the counterfactual curve $h \mapsto P_{\mathfrak{g},h}$ at each point is determined jointly by the policy functional ν and the cost metric $\mathfrak{g}$. This should be compared to the compensated demand curve in consumer theory that is determined jointly by the utility function (consumer preferences) and prices. Our methodology thus provides a natural way to model counterfactual probability distributions for policy analysis purposes with a statistical functional that reflects policymaker preferences and a metric over alternative distributions that reflects the cost of counterfactual distributions. On the other hand, the counterfactual objects defined in this paper have good statistical properties. Because we specify counterfactual distributions with local optimality conditions, our non-local

[2]All counterfactuals in this paper are descriptions of hypothetical states of nature. The main objective here is to develop a general framework to describe nonparametric counterfactual distributions with changes in scalar functionals. The conditions for causal interpretations of these counterfactuals are outside of the scope of this paper. But see Section 1.4.3 for an illustration of how the methodology of this paper can be combined with instrumental variables approach to causal inference.

counterfactuals depend smoothly on the status quo distribution P_0. Consequently, the effects in eq. (0.2) are not only identifiable, but are also consistently estimable with data generated from P_0, which requires smooth dependence on P_0 [see 49].

Literature Our methodology generalizes seemingly unrelated results of [36, 6, 7], and [30]. The former are approximate effects of changes in the distribution along *the hardest submodel* for a descriptive statistic ν on a structural estimand ψ obtained via the asymptotic distribution of a joint estimator $(\widehat{\nu}, \widehat{\psi})$. The latter are approximate Oaxaca-Blinder effects of changes in the means of covariates X on the quantiles of the outcome Y obtained via location shifts of the distribution of covariates. The results of [36, 6, 7] map into our framework by setting the cost parameter $\mathfrak{g}$ to the information metric[3]. The results of [30] map into our framework by setting the cost parameter $\mathfrak{g}$ to the L^2-Wasserstein metric[4]. We extend both sets of results by allowing counterfactual analysis with a general policy functional ν and cost metric $\mathfrak{g}$, and by describing changes of a non-infinitesimal magnitude h in both the entire distribution and a general outcome functional ψ of the distribution. Our derivation of the Wasserstein gradient is related to [70]. Our derivation of the influence function of the effect functional (0.2) is related to [41, 73]. References to technical literature are made in the appropriate sections of the paper.

Roadmap To comment on the technical aspects of the paper, Problems 1 and 2 are solved by applying von Mises calculus of infinitesimal changes in the statistical functionals ν and ψ. This is done by working with *influence functions* (scores) to compute the infinitesimal changes and integrating[5] the infinitesimal changes to obtain distributional and scalar counterfactuals of a non-infinitesimal magnitude h. Influence functions $\widetilde{\nu}_P, \widetilde{\psi}_P : \mathcal{X} \to \mathbb{R}$ are an important tool in the theory of asymptotic inference for estimators of the parameters $\nu(P), \psi(P)$ with data generated from P. The influence function obtains its name from the fact that the contribution of an additional observation X_n on the value of any regular estimator $\widehat{\psi}$ of parameter ψ at distribution P is $\widetilde{\psi}_P(X_n)/n$ in large samples [see e.g. 55, 56]. Influence functions are known for many estimators because they determine their robustness properties [e.g. 40, 44], asymptotic distribution [e.g. 56, 45] and efficiency bounds [e.g. 13, 65]. The influence function is a useful object for many purposes in econometrics [see 45] and can be computed with techniques of Ichimura and Newey (2017) [45].

Problem 2 is solved readily by using influence functions $\widetilde{\psi}_P$ to compute the infinitesimal changes in ψ along any regular path from P_0 to P_h and integrating them via the fundamental theorem of calculus. In the setting of Oaxaca-Blinder decompositions, this yields a computationally efficient method to evaluate the counterfactual effect in an arbitrary functional ψ of the distribution of the outcome variable Y without requiring access to the entire conditional distribution of Y given X. Our calculation extends the RIF regression technique of Firpo, Fortin, and Lemieux (2009) [30] to evaluate the change in ψ with arbitrary precision with a sequence of regressions, rather than only to first order with a single regression.

Problem 1 is solved in two installments: local (infinitesimal) analysis and global (non-infinitesimal) analysis. Locally at every distribution P, we solve for the gradient score with respect to the given metric $\mathfrak{g}$ of a functional ν in terms of the influence function $\widetilde{\nu}_P$. The influence function is the gradient score for the Fisher-Rao information metric. We discuss gradient

[3]The information metric, also known as the Fisher-Rao metric, is the metric tensor of the Bhattacharyya geodesic distance on the unit L^2 sphere. It is topologically equivalent to the Hellinger norm and the total variation distance. It coincides with the Kullback-Leibler divergence locally to second order. All these distances measure statistical dissimilarity of probability distributions.

[4]The Wasserstein metric, also known as the Kantorovich or Kantorovich-Rubinstein metric, measures the optimal cost in the Monge-Kantorovich problem of transporting population mass from one distribution to another.

[5]Integrating the gradient flow differential equation on the space of probability distributions $\mathcal{P}$ endowed with a manifold structure to compute distributional changes and applying the fundamental theorem of calculus to compute scalar changes along the integral paths of distributions.

flow counterfactuals of the information and 2-Wasserstein metrics. We also generalize these by allowing for cost-distribution weights on the sample space $\mathcal{X}$ in both metrics, and provide formulas that map the influence function into the gradient score. For completeness, we also provide a formula for the influence function using approximations to the identity (i.e. kernels) in the Riesz representation of the pathwise derivative following [45]. From the local analysis we obtain a *score field*, which consists of a local perturbation at each counterfactual distribution $P \in \mathcal{P}$. To connect the infinitesimal perturbations at different distributions $P \in \mathcal{P}$, we use the structure of a manifold on $\mathcal{P}$, modeled on exponential Orlicz spaces following [61]. To obtain the path of counterfactual distributions $P_{\mathfrak{g},\nu,h,P_0}$, we solve the resulting gradient flow differential equation on $\mathcal{P}$ with the initial condition P_0. We then use the smoothness properties of the differential equation on the initial condition P_0 to derive the influence function of the scalar effect $\Delta_{\mathfrak{g},\nu,h}\psi(P_0)$ in (0.2) along the integral curve $P_{\mathfrak{g},\nu,h,P_0}$ of the flow.

The gradient is the direction of least costly counterfactual states of nature $P_{\mathfrak{g},\nu,h,P_0}$ for the policy that increases parameter ν and incurs a cost equal to the distance $\mathfrak{g}(P_h, P_0)$ between the counterfactual state and the status quo state. The magnitude of the gradient score depends on the metric. The counterfactual distribution $P_{\mathfrak{g},\nu,h,P_0}$ is obtained by integrating the infinitesimal perturbations (gradient scores) starting from P_0 until the increment h in ν is attained. This requires a rescaling of the gradient scores such that the local change in the policy functional is always unitary. The effect of ν on ψ is the change in ψ along the curve of counterfactual distributions parametrized by the change h in ν. Results obtained in this paper extend those of Andrews, Gentzkow, and Shapiro (2017) in [6] and those of Firpo, Fortin, and Lemieux (2009) in [30]. [6] describe the infinitesimal change in a structural estimand ψ along the influence function $\widetilde{\nu}_P$ of a descriptive statistic ν by working with the asymptotic covariance[6] of a joint estimator $(\widehat{\psi}, \widehat{\nu})$ and, implicitly, with information gradient flows. [30] describe the infinitesimal change in quantiles of Y along the Wasserstein gradient of the mean of X by working with location shifts of the distribution of X and, implicitly, with optimal transport flows.

Bonus As an application of our methodology, we obtain a nonparametric Oaxaca-Blinder counterfactual composition interpretation of the OLS and IV estimators. From our methodology, it follows that there exist 1-dimensional models of counterfactual distributions parametrized by (the change in) the mean of the covariate X, such that (the large sample limit of) the estimator measures the local effects on the mean of the outcome variable Y in these models. Contrary to a common belief, the nonparametric counterfactual distributions of covariates associated with the OLS coefficient are *not* location shifts in X, if the status quo distribution of X is not Gaussian. The IV coefficient measures the effect on the mean of Y of the variation in the mean of X induced by changing the marginal distribution of the instrument Z. By contrast with OLS , the counterfactual distributions associated with the IV coefficient do not preserve the conditional distributions of Y given X in general. Similarly with OLS , the IV counterfactual distributions are not location shifts in either X or Z, if the joint status quo distribution of (Y, X, Z) is not Gaussian. Moreover, location shifts in the distribution of covariates X are a special case of transport counterfactuals that correspond to gradient flows in the Wasserstein metric. Their infinitesimal scalar effects take the form of an average derivative, which is a very common estimand in semiparametric estimation.

Technical contribution Nonparametric influence function of the GMM functional, information gradient flow characterization of the hardest submodel, Lebesgue differentiation formula for the influence function, Wasserstein gradient formula, Lipschitz condition for existence of integral curves, influence function of the global OLS effect.

[6] Asymptotic covariance of a joint estimator $(\widehat{\psi}, \widehat{\nu})$ is the inner product $\mathrm{E}_P[\widetilde{\psi}_P \widetilde{\nu}_P]$ of influence functions $\widetilde{\psi}_P$ and $\widetilde{\nu}_P$ and has the value of the derivative of ψ in the direction of the information gradient $\widetilde{\nu}_P$ of ν.

The rest of this paper is organized as follows. A nontechnical Section 1 presents two high-level results that cover Problems 1 and 2 while relying on minimal amount of notation and contains several detailed examples that illustrate our methodology with common econometric estimators: ATE , OLS , IV and GMM . Section 2 relies on standard notation of semiparametric efficiency theory to derive gradient scores from the influence function, characterizes the influence function and the counterfactual distributions and effects along the influence function that can be obtained from the asymptotic distribution of regular estimators as suggested by Gentzkow and Shapiro (2015) [36]. Section 3 uses the exponential manifold structure of [61] to prove existence of gradient flows for the functionals discussed in the sequel and derives the influence function of their scalar effects (0.2).

0.1 Notation, conventions, preliminaries.

$(\mathcal{X}, \mathcal{A})$	Sample space, with $\mathcal{X} \subset \mathbb{R}^d$ and Borel σ-algebra $\mathcal{A}$;
$\mathcal{F}, \mathcal{P}$	Collection of counterfactual probability distributions on $(\mathcal{X}, \mathcal{A})$;
$F_0, P_0,$	Status quo probability distributions in $\mathcal{P}$;
F_h, P_h	Counterfactual probability distributions, path of counterfactual distributions;
F_X	Cumulative distribution function of random variable X;
f_X, f, ϱ, p, q	Probability density functions of random variable X, probability measure P;
$\mathcal{T}_P$	Tangent space at point $P \in \mathcal{P}$;
u, v_P, w_P	Score functions $v_P = \frac{d}{dt} \log f_t$, i.e. tangent vectors in $\mathcal{T}_P$;
ν, ψ	Statistical functionals (also parameters) $\nu, \psi : \mathcal{P} \to \mathbb{R}$. E.g. descriptive statistics or large-sample limits of estimators $\widehat{\nu}, \widehat{\psi}$;
$\widetilde{\psi}_P, \widetilde{\nu}_P$	Influence functions of statistical functionals at probability distribution $P \in \mathcal{P}$;
$\Delta_{\nu,h}\psi$	Counterfactual effect of increasing ν by h on ψ;
$P_{\mathfrak{g},\nu,h}$	Gradient flow path of counterfactual distributions for functional ν in metric $\mathfrak{g}$;
$\Delta_{\mathfrak{g},\nu,h}\psi$	Counterfactual effect on ψ of increasing ν by h along the gradient flow path in metric $\mathfrak{g}$, i.e. $\Delta_{\mathfrak{g},\nu,h}\psi = \psi(P_{\mathfrak{g},\nu,h}) - \psi(P_0)$;
$\mathfrak{g}, \mathfrak{g}_P$	Metric on $\mathcal{P}$, i.e. collection of inner-products $\mathfrak{g}_P(\cdot,\cdot)$ on tangent spaces $\mathcal{T}_P$ of model $\mathcal{P}$, and the induced geodesic distance function $\mathfrak{g}(\cdot,\cdot)$ on $\mathcal{P}$;
$\mathfrak{g}_{\mathbf{F}}, \mathfrak{g}_{\mathbf{K}}$	Fisher-Rao information metric, 2-Wasserstein-Kantorovich metric;
$\nabla_{\mathfrak{g}}\psi_P$	Gradient score function in $\mathcal{T}_P$ of functional ψ with respect to metric $\mathfrak{g}$;
$\boldsymbol{v}, \boldsymbol{w}$	Velocity vector fields in $\overline{\{\nabla_x\varphi\ ;\ \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(P;\mathbb{R}^d)}$;
δ_x	Point mass at $x \in \mathcal{X}$;
$\mathbb{1}_A$	Indicator function of set A;
$L^2(P)$	Space of measurable maps $f : \mathcal{X} \to \mathbb{R}$ with $\int_{\mathcal{X}} \lvert f\rvert^2 dP < +\infty$;
$L_0^2(P)$	Subspace of maps $f \in L^2(P)$ with zero P-mean;
$L^2(P;\mathbb{R}^d)$	Space of measurable maps $f : \mathcal{X} \to \mathbb{R}^d$ with $\int_{\mathcal{X}} \lVert f\rVert_{\mathbb{R}^d}^2 < +\infty$;
J	Open interval $J \subset \mathbb{R}$ containing $0 \in \mathbb{R}$;
$\mathscr{L}^d$	Lebesgue measure on $\mathbb{R}^d$;
μ, χ, ζ	means of the outcome, covariate, and instrument random variables Y, X, Z;
t, s	time parameter of unscaled flow curves;
h	time parameter of scaled flow curves;

1. Scalar counterfactuals

We begin with a warm up exercise by solving Problem 2 in the setting of Oaxaca-Blinder (OB) decompositions in a novel way. This result for OB decompositions is then used to explain our approach to counterfactual distributional analysis with a policy functional by relying on minimal amount of notation. Empirical implications of our methodology are then illustrated with several examples of counterfactual distributions and effects implied by common estimators.

1.1 Oaxaca-Blinder setting. Let $F_{XY,0} : \mathcal{X} \times \mathbb{R} \to [0,1]$ denote the joint cumulative distribution function of an outcome random variable Y and covariates X, also called the status quo distribution and also denoted by F_0. The goal of OB-type counterfactual analysis is to evaluate the effect of a change in the distribution of covariates X on the marginal distribution of the outcome Y. Let $F_{X,1}$ denote the counterfactual distribution of covariates, e.g. formulated by analyzing a policy that transforms X or estimated with a sample drawn from a control population of agents. The conditional distribution $F_{Y|X}(\cdot|x) = \int_{-\infty}^{\cdot} f_{XY,0}(x,y)/f_X(x)\,\mathrm{d}y$ describes the structure of stochastic assignment of outcomes y to agents with covariates x. Under the core OB assumption that the outcome assignment $F_{Y|X}$ is unaffected by the change in the distribution of covariates, the counterfactual and the status quo marginal distributions of the outcome are given by

$$F_{Y,h}(y) = \int_{\mathcal{X}} F_{Y|X}(y|x)\,\mathrm{d}F_{X,h}(x), \qquad h = 1, 0. \tag{1.1}$$

To quantify the difference between $F_{Y,1}$ and $F_{Y,0}$, empirical researchers often look at counterfactual effects on scalar parameters of the outcome distribution. For example, Oaxaca (1973) [58] and Blinder (1973) [14] studied the difference in the expected values of $F_{Y,1}$ and $F_{Y,0}$.

For a given functional

$$\psi : \mathcal{F}_Y \to \mathbb{R}$$

of the marginal distribution $F_Y \in \mathcal{F}_Y$, the effect of a small perturbation in the distribution of Y on ψ can be approximated by the von Mises (1947) [53] formula:

$$\tfrac{d}{dh}_{|h=0}\psi(F_{Y,h}) = \lim_{h\to 0} h^{-1}[\psi(F_{Y,h}) - \psi(F_{Y,0})] = \int_{\mathbb{R}} \widetilde{\psi}_{F_{Y,0}}(y)\,\mathrm{d}[F_{Y,1} - F_{Y,0}](y) \tag{1.2}$$

where $F_{Y,h} = (1-h)F_{Y,0} + hF_{Y,1}$ is the mixture model interpolating the status quo and the counterfactual distributions of Y, and $\widetilde{\psi}_{F_{Y,0}} : \mathbb{R} \to \mathbb{R}$ is the influence function of parameter ψ. A rigorous definition of the influence function[7] is important to the results presented later in this paper but requires investment in notation and is therefore postponed to Section 2. Any functional that can be estimated at the parametric $\sqrt{n}$-rate with some uniformity in the asymptotic distributions has an influence function. Influence functions for all functionals discussed in the sequel are provided. By definition, the influence function $\widetilde{\psi}_{F_Y}$ has finite second moment and zero expectation under F_Y. From eq. (1.2) it follows that $\widetilde{\psi}_{F_{Y,0}}(y) = \frac{d}{dh}_{|h=0}\psi(F_{Y,h})$ when we take $F_{Y,1}$ to be the point mass at y. This suggests the interpretation of $\widetilde{\psi}_{F_Y}(y)$ as the effect of replacing an infinitesimal part of the distribution F_Y with a point mass at y on the parameter $\psi(F_Y)$. For example, the influence function of the mean functional

$$\mu(F_Y) = \int_{\mathbb{R}} y\,\mathrm{d}F_Y(y) \qquad \text{is} \qquad \widetilde{\mu}_{F_Y}(y) = y - \mu(F_Y). \tag{1.3}$$

Replacing part of the distribution F_Y with mass in the tail of the distribution has a larger effect on the mean than replacing part of F_Y with mass close to the expected value.

[7] Riesz representation of the derivative of the map ψ with respect to the information metric inner product [47].

Following Firpo, Fortin, and Lemieux (2009) [30], by using the linearity of expectations, the law of iterated expectations and the core OB assumption that the conditional distributions $F_{Y|X}$ are unaffected by changes in the law of X, we obtain from eqs. (1.1) and (1.2):

$$\begin{aligned}\tfrac{d}{dh}|_{h=0}\psi(F_{Y,h}) &= \int_{\mathbb{R}} \widetilde{\psi}_{F_{Y,0}}(y)\, \mathrm{d}F_{Y,1}(y) \\ &= \int_{\mathcal{X}}\int_{\mathbb{R}} \widetilde{\psi}_{F_{Y,0}}(y)\, \mathrm{d}F_{Y|X}(y|x)\, \mathrm{d}F_{X,1}(x) \\ &= \int_{\mathcal{X}} \mathrm{E}\Big[\widetilde{\psi}_{F_{Y,0}}(Y)|X=x\Big]\, \mathrm{d}F_{X,1}(x). \end{aligned} \tag{1.4}$$

Equation (1.4) shows that the approximate effect of evolving the distribution of covariates $F_{X,0}$ toward $F_{X,1}$ on parameter $\psi(F_{Y,0})$ can be computed by averaging the nonparametric regression of the transformation $\widetilde{\psi}_{F_{Y,0}}(Y)$ of the outcome variable Y on X with respect to the counterfactual distribution of covariates. Furthermore, the exact counterfactual effect of evolving $F_{X,0}$ to $F_{X,1}$ can be computed by integrating the infinitesimal changes $\frac{d}{dh}\psi(F_{Y,h})$ along the path $h \mapsto F_{Y,h}$ via the fundamental theorem of calculus:

Proposition 1 (Effect of a change in the distribution of X on a statistic of Y)**.** Suppose the marginal distribution of covariates can be changed from $F_{X,0}$ to $F_{X,1}$ without affecting the conditional distributions $F_{Y|X}$. Then the effect on the statistic ψ of F_Y can be computed as

$$\Delta\psi = \psi(F_{Y,1}) - \psi(F_{Y,0}) = \int_{[0,1]}\int_{\mathcal{X}} \mathrm{E}\Big[\widetilde{\psi}_{F_{Y,h}}(Y)|X=x\Big]\, \mathrm{d}F_{X,1}(x)\, \mathrm{d}h \tag{1.5}$$

where $F_{Y,h} = (1-h)F_{Y,0} + hF_{Y,1}$ with the counterfactual distribution $F_{Y,1}$ given in eq. (1.1), and $\widetilde{\psi}_F$ is the influence function of parameter ψ at distribution F.

Proposition 1 extends the result of Firpo, Fortin, and Lemieux (2009) [30] who obtained a first-order approximation to $\Delta\psi$ with (1.4). Proposition 1 can be used to estimate the exact counterfactual effect on $\psi(F_Y)$ without estimating the entire conditional distribution $F_{Y|X}$ in order to find the counterfactual distribution $F_{Y,1}$ via eq. (1.1) as described in Chernozhukov, Fernández-Val, and Melly (2013) [20]. For example, in the case of the quantile functional $\psi_\tau(F_Y) = F_Y^{-1}(\tau)$, it is intuitively clear that knowing only a small part of the counterfactual distribution function $F_{Y,1}$ in a neighborhood of the counterfactual quantile $\psi_\tau(F_{Y,1})$ is required to find the effect $\Delta\psi_\tau$. Therefore, knowing only a small subset of the conditional probabilities $\{F_{Y|X}(y|x)\ ;\ \psi_\tau(F_{Y,1}) - \epsilon < y < \psi_\tau(F_{Y,1}) + \epsilon\}$ is required to find $\Delta\psi_\tau$. This intuition is made precise with eq. (1.5) by substituting the influence function of the quantile

$$\psi_\tau(F_Y) = F_Y^{-1}(\tau), \qquad \widetilde{\psi}_{\tau,F}(y) = [\tau - \mathbb{1}\{y \le \psi_{\tau,F}\}]/f_Y(\psi_{\tau,F}) \tag{1.6}$$

and noting that the counterfactual effect $\Delta\psi_\tau$ depends on the conditional distribution $F_{Y|X}$, through the regression of indicators $\mathbb{1}\{Y \le \psi_\tau(F_{Y,h})\}$ on X and density values $f_{Y,h}(\psi_\tau(F_{Y,h}))$, only at the τ-quantiles on the path $F_{Y,h}$. Moreover, eq. (1.5) also suggests an iterative procedure to evaluate $\Delta\psi_\tau$ with regression and density estimation methods, leading to a new estimator of the unconditional quantile effect.

1.2 General nonparametric setting. We now extend the Oaxaca-Blinder setting to allow counterfactual analysis of a scalar parameter along a general and, furthermore, implicit path of joint counterfactual distributions F_h as follows. A more suitable for our purposes form of the von Mises formula for the pathwise derivative (1.2) is the *inner product* (i.e. covariance)

$$\tfrac{d}{dh}|_{h=0}\psi(F_h) = \int \widetilde{\psi}_{F_0}(x,y) v_{F_0}(x,y)\, \mathrm{d}F_0(x,y) \tag{1.7}$$

of the influence function $\widetilde{\psi}_F$ of the parameter ψ and the *score function* (i.e. derivative of log likelihood)

$$v_{F_0}(x,y) = \tfrac{d}{dh}_{|h=0} \log f_h(x,y) = \dot{f}_0(x,y)/f_0(x,y) \qquad x,y \in \mathcal{X} \times \mathbb{R} \tag{1.8}$$

of the 1-dimensional parametric model F_h, obtained by smoothly interpolating the joint status quo distribution F_0 and the joint counterfactual distribution F_1.[8]

The score v_{F_0} is the infinitesimal perturbation (i.e. direction and magnitude of change) in the status quo distribution F_0 along the path $h \mapsto F_h$. The influence function $\widetilde{\psi}_{F_0}$ is a particular score associated to the functional ψ by the covariance inner product (the information metric). Specifically, $\widetilde{\psi}_{F_0}$ is the score that maximizes the covariance inner product, as follows readily from the Cauchy-Schwarz inequality. Formula (1.7) can be used to find approximate and exact counterfactual changes in ψ along *any* regular parametric model by taking its scores and computing the covariance with the influence functions of ψ as in Proposition 1. The main point this paper makes is that the pathwise derivative (1.7) can be used to make the path of counterfactual distributions F_h *implicit* by replacing it with a *collection of scores* $\{v_F\}_{F\in\mathcal{F}}$, one at every hypothetical distribution $F \in \mathcal{F}$. Furthermore, this collection of scores can be specified with counterfactual local changes in a *policy functional* $\nu : \mathcal{F} \to \mathbb{R}$ that are often discussed by practitioners informally [see 36, 6, 7, references therein, citations thereof, and Section 1.5].

Proposition 2 (Path of counterfactual distributions of a score field, effect on a statistic)**.** Suppose a scalar parameter ν defined on $\mathcal{F}$ can be changed locally at each counterfactual distribution $F \in \mathcal{F}$ along the score v_F. Suppose the collection of scores $\{v_F\}_{F\in\mathcal{F}}$ satisfies a local Lipschitz condition. Then at any status quo point $F_0 \in \mathcal{F}$, there exists a unique path of counterfactual distributions $\{F_h\}_{h\in J}$ defined on an open interval J containing 0, such that $\nu(F_h) = \nu(F_0) + h$ is the parameter of the path, and its score $\frac{d}{dh} \log f_h$ is proportional to v_{F_h} for every $h \in J$.

Moreover, the effect of increasing the policy parameter ν by $h \in J$ on another scalar parameter ψ defined on $\mathcal{F}$ is given by

$$\Delta_{\nu,h}\psi = \psi(F_h) - \psi(F_0) = \int_{[0,h]} \int \widetilde{\psi}_{F_t} \left[v_{F_t} / \mathrm{Cov}_{F_t}(\widetilde{\nu}_{F_t}, v_{F_t}) \right] \mathrm{d}F_t \, \mathrm{d}t, \tag{1.9}$$

where $\widetilde{\psi}_{F_t}$ is the influence function of ψ, and $\widetilde{\nu}_{F_t}$ is the influence function of ν.

Proposition 2 says that counterfactual distributions and the corresponding scalar effects can be specified implicitly by considering infinitesimal changes in a policy parameter $\nu(F) \in \mathbb{R}$ and the infinitesimal perturbations (scores) v_F to the distribution F that generate these local changes in ν. Under a Lipschitz condition, the scores can be integrated into a curve of counterfactual distributions. The main difficulty in this result is that scores at different distributions are not readily comparable because they belong to different tangent spaces. This is because $\mathcal{F}$ is not a linear space. The technical solution is to parametrize $\mathcal{F}$ by a normed linear space in a smooth and consistent way so that all tangent spaces can be identified with that single linear space. Such parametrization connects scores at different distributions and also makes the Lipschitz condition precise. The formula for the scalar effects follows by the fundamental theorem of calculus.

In order to apply Proposition 2, a score field $\{v_F\}_{F\in\mathcal{F}}$ is required. These scores can be found in the empirical interpretation of an estimator (ATE , OLS , IV), and in the asymptotic distribution of an estimator. In the rest of the paper we will discuss a nonparametric framework and results that obtain the scores v_F and counterfactual distributions F_h from the combination

[8] The formula is written for a general statistic ψ of the joint distribution F_{YX} and a general regular path F_h from the status quo distribution F_0 to the counterfactual distribution F_1; in the case where ψ is a statistic of the marginal distribution F_Y, the influence function $\widetilde{\psi}_F$ depends only on y; in the case where F_h has constant conditional distributions $F_{Y|X}$ as in Proposition 1, the score v depends only on x.

of a functional ν and a metric of distance or cost $\mathfrak{g}$ on $\mathcal{F}$ by solving local optimality conditions. Specifically, we will take v_F to be the direction of most rapid change in ν. These locally optimal counterfactual distributions and scores will be denoted by $F_{\mathfrak{g},\nu,h}$ and $\nabla_{\mathfrak{g}}\nu_F = \frac{d}{dh}\log f_{\mathfrak{g},\nu,h}$, and called the $\mathfrak{g}$-flow curves and $\mathfrak{g}$-gradients of the functional ν. The metric $\mathfrak{g}$ is a generalization of the covariance inner product (1.7) to allow policy changes in ν that are optimal in the sense most suitable to the economic application at hand rather than in the sense of statistical information distance.

We provide the technical details of the local analysis of gradient scores in Section 2 and of the global (non-infinitesimal) analysis of integral curves of score fields in Section 3. In the rest of Section 1 we showcase our framework by examining nonparametric counterfactual interpretation of several common estimators through the lens of the von Mises derivative (1.7) and Proposition 2.

1.3 Effect of changing the mean of an indicator X on Y: ATE. A simple example that can be studied with either Proposition 1 or 2 is the potential outcomes framework. Suppose we want to estimate the Oaxaca-Blinder effect of increasing the proportion of unionized workers χ on a distributional statistic ψ of wages Y:

$$\tfrac{d}{d\chi}\psi(F_Y), \quad \text{where} \quad \psi : \mathcal{F}_Y \to \mathbb{R}, \quad \chi = \mathrm{E}_F[X],$$

and X is the indicator of union status. Let $f_{X,\chi} = (1-\chi)\delta_0 + \chi\delta_1$ denote the density function of a hypothetical distribution of X, where δ_x is the point mass at x. Under the OB condition that changes in the mean parameter χ do not affect the conditional distributions $\{F_{Y|X=0}, F_{Y|X=1}\}$, we obtain a path of joint counterfactual distributions $\{F_{XY,\chi}\}_{0<\chi<1}$ i.e. a 1-dimensional parametric model. The score function of this path is

$$\begin{aligned} v_F(x) &= \tfrac{d}{d\chi}\log f_{Y|X}(y|x) f_{X,\chi}(x) = \frac{-1}{1-\chi}\delta_0(x) + \frac{1}{\chi}\delta_1(x) \\ &= \frac{x-\chi}{\chi(1-\chi)} = \frac{\widetilde{\chi}_F(x)}{\mathrm{Var}_F[X]} \qquad \text{a.s. } \delta_{\{0,1\}}, \end{aligned}$$

where

$$\chi(F) = \int x \,\mathrm{d}F, \qquad \widetilde{\chi}_F(x) = x - \chi(F) \tag{1.10}$$

is the influence function of the mean functional of the joint distribution F. We want to compute the infinitesimal effect of increasing the proportion of unionized workers χ on a statistic ψ. By the von Mises formula (1.7), this can be computed as the covariance between the influence function of the statistic $\widetilde{\psi}_F$ and the score function of the counterfactual model $v_F = \widetilde{\chi}_F / \mathrm{Var}_F[X]$, which happens to be the normalized influence function of the mean parameter χ of our model:

$$\tfrac{d}{d\chi}\psi(F_\chi) = \int \widetilde{\psi}_F(y)\left[\widetilde{\chi}_F(x)/\mathrm{Var}_F[X]\right] \mathrm{d}F(x,y) = \mathrm{E}[\widetilde{\psi}_F(Y)|X=1] - \mathrm{E}[\widetilde{\psi}_F(Y)|X=0].$$

If we take ψ to be the mean μ of Y defined in (1.3), we obtain the population level counterfactual interpretation of the average treatment effect $\frac{d}{d\chi}\mu(F_\chi) = \mathrm{E}[Y|X=1] - \mathrm{E}[Y|X=0] = \textsc{ate}$.

The point we make with this example is that the influence function of a statistical functional serves two different purposes in the present paper: (i) $\widetilde{\psi}_F$ is a computational device for the pathwise derivative of the outcome parameter ψ; (ii) $\widetilde{\chi}_F$ is a direction of perturbation to the distribution F that controls the counterfactual value of the policy parameter χ. Although we started with an explicit path of counterfactual distributions, Proposition 2 says that we can also start with the local perturbations $\widetilde{\chi}_F$ and local effects $\frac{d}{d\chi}\psi$ and add them together.

In the case of an indicator policy variable, there is only one possible path of marginal counterfactual distributions, therefore a unique way to change the mean χ, and consequently a unique

OB-effect $\frac{d}{d\chi}\psi(F_Y)$ on an outcome functional ψ. However, as discussed in the Introduction, the same question has been raised by practitioners in the case of a general random element X and a general policy functional ν of its distribution. The latter question is ill-posed; and counterfactual distributions generated by the score field of influence functions $\widetilde{\nu}_F$ is one possibility among many. A new methodology is therefore required to provide practitioners with tools to model and compute infinitesimal effects $\frac{d}{d\nu}\psi(F)$ and non-infinitesimal effects $\Delta_{\nu,h}\psi(F)$. This is achieved by finding natural perturbations (scores) $v_{\nu,F}$ to F that increase ν infinitesimally, and counterfactual distributions $h \mapsto F_{\nu,h}$ parametrized by the change h in ν.

1.4 Effects of changing the mean of a scalar X on Y: OLS, IV, average derivative. We now consider three different examples of Oaxaca-Blinder counterfactual composition effects by changing the mean of a general scalar covariate X. The purpose of these examples is to illustrate that: (i) a nonparametric interpretation in terms of counterfactual distributions is available for common structural estimators; (ii) there are many natural ways to change the probability distribution in order to control a scalar statistical parameter (the mean here). To quantify the distributional changes, we consider their effects on the mean and the τ-quantile of the outcome variable Y, and let χ, μ and ψ_τ denote these functionals in the sequel:

$$\chi(F) = \int x \, \mathrm{d}F, \quad \mu(F) = \int y \, \mathrm{d}F, \quad \psi_\tau(F) = F_Y^{-1}(\tau).$$

We reserve notation ν and ψ for generic policy and outcome functionals.

1.4.1 **Linear regression (OLS)** Fortin, Lemieux, and Firpo (2011) [33, p. 7] write:

> "the coefficient β in a standard regression ... $\mathrm{E}(Y|X) = \beta_0 + \beta X$ ... can be interpreted as the effect of increasing the mean value of X on the mean value of Y ... using the law of iterated expectations, $\mathrm{E}(Y) = \beta_0 + \beta \mathrm{E}(X)$".

In the structural equation model with a linear conditional expectation function, *any* change in the distribution of covariates X to a counterfactual $F_{X,h}$ that has $\chi(F_{X,h}) = \chi(F_{X,0}) + h$ and satisfies the Oaxaca-Blinder condition that $F_{Y|X}$ are unaffected, generates the same linear effect $\Delta_{\chi,h}\mu = \beta h$ on the mean of outcome Y. While we may not think of the linear model as the exact description of the data, it is commonly used in empirical work, and the OLS coefficient β is commonly interpreted in this way.

However, this interpretation of the regression coefficient β is also valid in the completely nonparametric (nonlinear) model as well, with the caveat that we must be precise about the counterfactual distributions $F_{X,h}$ and consider *infinitesimal* changes. The caveat arises because the least squares projection (β_0, β) of Y on the linear span of $\{1, X\}$ depends not only on the conditional distributions $F_{Y|X}$, but also on the marginal distribution F_X [74]. For a small perturbation in F_X, changes in parameters $\mu, \chi, \beta_0, \beta : \mathcal{F} \to \mathbb{R}$ are approximately linear by smoothness. Furthermore, these small changes are related by the chain rule through the regression equation:

$$\Delta\mu = \Delta\beta_0 + (\Delta\beta)\chi + \beta(\Delta\chi) + o(\Delta F)$$

which is in general different from $\beta(\Delta\chi)$. This means that the OLS coefficient β measures the effect of changes in χ on μ only for a particular perturbation in F_X. Somewhat surprisingly, this perturbation is *not* a location shift $X \mapsto X + h$, if the distribution of X is not Gaussian.

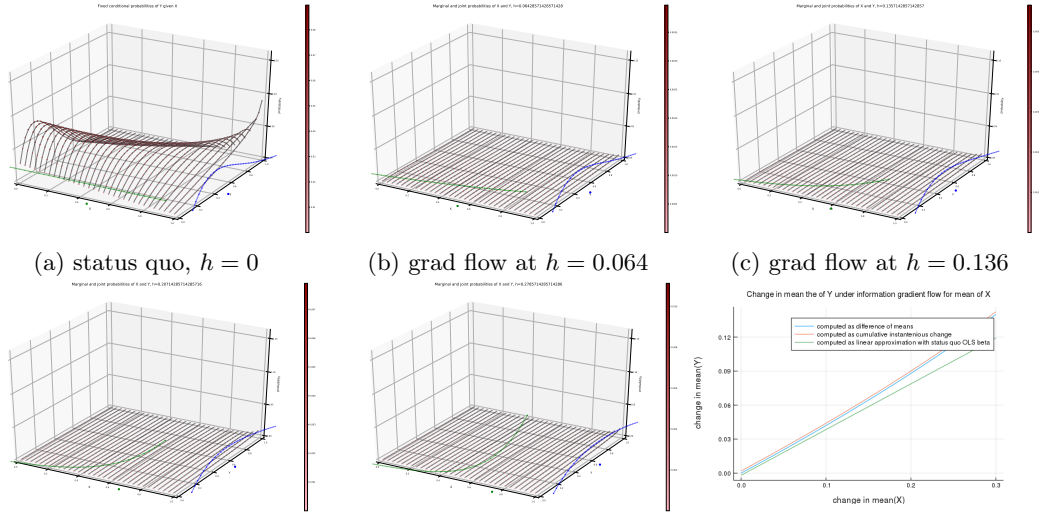
To see the score $v_{\mathrm{OLS},F}$ with the infinitesimal effect $\beta(F) \cdot \Delta\chi$ on $\mu(F)$, recall the OLS coefficient formula:

$$\beta(F) = \frac{\mathrm{Cov}_F(Y,X)}{\mathrm{Var}_F(X)} = \frac{\int \big[y - \mu(F)\big]\big[x - \chi(F)\big]\mathrm{d}F}{\int \big[x - \chi(F)\big]^2 \mathrm{d}F} = \int \widetilde{\mu}_F(y) \frac{\widetilde{\chi}_F(x)}{\mathrm{Var}_F[X]} \, \mathrm{d}F(x,y) \qquad (1.11)$$

where we unpacked the expectation operators in $\mathrm{Cov}(Y, X)$ and $\mathrm{Var}(X)$, substituted our functional notation μ and χ for the means of Y and X, and recognized the influence functions of these functionals. Comparing the last expression in (1.11) with the von Mises formula (1.7), we see that β is the *derivative* of the mean functional μ of Y along the rescaled influence function $v_{\mathrm{OLS},F} = \widetilde{\chi}_F / \mathrm{Var}_F[X]$ of the mean χ of X. Furthermore, applying Proposition 2 with the score field $\{v_{\mathrm{OLS},F}\}_{F\in\mathcal{F}}$, we conclude that for every status quo distribution F_0 with enough regularity, there is a unique parametric model $\{F_{\mathrm{OLS},h}\}_{h\in J}$ of counterfactual distributions that has these scores. The parameter of this model is the change $h = \chi(F_{\mathrm{OLS},h}) - \chi(F_0)$ in the mean of X, and the infinitesimal effects on the mean μ of Y of these counterfactual distributions are given by the OLS coefficients:

$$\tfrac{d}{dh}\mu(F_{\mathrm{OLS},h}) = \beta(F_{\mathrm{OLS},h}) \qquad \text{every } h \in J. \tag{1.12}$$

In other words, if we interpret the linear regression coefficient β as the effect of increasing the mean of X on the mean of Y and allow general misspecification of the population distribution $F_0 \in \mathcal{F}$, then $F_{\mathrm{OLS},h}$ is the (unique in a suitable sense) model of counterfactual distributions that are consistent with our interpretation of the functional β on $\mathcal{F}$.



(a) status quo, $h = 0$ (b) grad flow at $h = 0.064$ (c) grad flow at $h = 0.136$

(d) grad flow at $h = 0.207$ (e) grad flow at $h = 0.279$ (f) change in the mean μ of Y

Figure 1: OLS counterfactual distributions

Furthermore, by eq. (1.9), the non-infinitesimal effect of a real change h in χ on μ is given by the path integral of OLS coefficients

$$\begin{aligned}\Delta_{\mathrm{OLS},h}\mu = \mu(F_{\mathrm{OLS},h}) - \mu(F_0) &= \int_0^h \beta(F_{\mathrm{OLS},t})\,\mathrm{d}t \\ &= \int_0^h \int_{\mathcal{X}} \Big[\mathrm{E}_{f_0}[Y|X](x) - \mu(F_t)\Big] \frac{x - \chi(F_t)}{\mathrm{Var}_{f_t}[X]} \frac{f_{X,t}}{f_{X,0}}(x)\,\mathrm{d}F_{X,0}(x)\,\mathrm{d}t.\end{aligned} \tag{1.13}$$

The path $h \mapsto F_{\mathrm{OLS},h}$ is the integral curve of the ordinary differential equation in the space of

counterfactual probability distributions $\mathcal{F}$:

$$f_{\text{OLS},h}(x,y) = f_{Y|X}(y|x) f_{X,h}(x) \tag{1.14}$$

$$f_{X,h} = \frac{e^{u(h)}}{\mathrm{E}_{f_0}[e^{u(h)}]} f_{X,0} \tag{1.15}$$

$$\dot{u}(h) = \frac{x - \mathrm{E}_{f_0}[X]}{\mathrm{Var}_{f_h}[X]}, \quad u(0) = 0. \tag{1.16}$$

The right-hand side of the differential equation (1.16) is given by the rescaled influence functions of the mean functional χ. The influence function $\widetilde{\chi}_F$ is the direction for changing the mean χ that has the most rapid (gradient) change with respect to the covariance (information) metric $\mathfrak{g}$ in eq. (1.7). The scaling is such that the magnitude of the local change in χ is always unitary, and the parameter of the integral curve is the non-infinitesimal change in χ. The equation is written in terms of the score functions u at the status quo distribution F_0, because scores form a linear space that is required by the theory of differential equations. The corresponding counterfactual probability distributions, which do not form a linear space, are obtained by transforming scores at f_0 into densities near f_0 via the exponential map (1.15). The geometric properties that are required by the theory of differential equations of this parameterization of the nonparametric model $\mathcal{F}$ by the scores have been worked out by Pistone and Sempi (1995) [61] who were apparently motivated by Efron (1975) [29]. The regularity condition to guarantee existence of a solution $\{F_{\text{OLS},h}\}_{h\in J}$ with this parameterization is that the distribution of X has a moment generating function at F_0.

The effect $\Delta_{\text{OLS},h}\mu$ can be estimated by discretizing the integral $\int_{[0,h]} \mathrm{d}t$ above and updating recursively the mean, variance and likelihood ratio of the empirical marginal distribution of X and the mean of the marginal distribution of Y, starting from the status quo values (i.e. forward Euler method).

We run Monte Carlo experiments that produce a graphical representation of the OLS curve, see Figure 1. We solve for the OLS curve analytically and for the influence function of the effect $\Delta_{\text{OLS},h}\mu$ in Section 3.

1.4.2 **Average derivative** Consider a closely related example. Firpo, Fortin, and Lemieux (2009) [30] consider the OB counterfactual composition effect of the location shift

$$f_{\text{LSH},h}(x,y) = f_{Y|X}(y|x) f_{X,0}(x-h) \tag{1.17}$$

in the distribution of covariates F_X, while holding the outcome assignment $F_{Y|X}$ fixed. They show that the infinitesimal change in the τ-quantile ψ_τ of the outcome variable Y is given by the average derivative of the conditional expectation of its influence function $\widetilde{\psi}_{\tau,F}$:

$$\alpha := \int \left[\tfrac{d}{dx} \int \frac{\tau - \mathbb{1}\{y \le \psi_\tau(F)\}}{f_Y(\psi_{\tau,F})} \, \mathrm{d}F_{Y|X}(y|x) \right] \mathrm{d}F_X(x) = \tfrac{d}{dh} \psi_\tau(F_{\text{LSH},h}). \tag{1.18}$$

Fortin, Lemieux, and Firpo (2011) [33, p. 8] describe this result as follows:

> "the estimated coefficient [α] can be interpreted as the effect of increasing the mean value [χ] of X on the unconditional quantile [ψ_τ of Y]".

We claim that it is a corollary to [30] that the curve of counterfactual distributions $F_{\text{OLS},h}$, defined implicitly by eqs. (1.14) to (1.16), is not the location shift (1.17) in the distribution of X, contrary to the common belief. This can be verified by either comparing α to the change in the quantile ψ_τ of Y along the OLS model $F_{\text{OLS},h}$, or by comparing β to the change in the mean μ of Y along the location shift model $F_{\text{LSH},h}$. Let us describe these.

The local effect of changing the mean χ of X with the OLS scores on the quantile ψ_τ of Y follows by the von Mises formula (1.7). It is given by the covariance of the influence function

$\widetilde{\psi}_{\tau,F}$ of the quantile functional (1.6) and the score $v_{\text{OLS},F}$ of the OLS path of counterfactual distributions (1.11):

$$\tfrac{d}{dh}\psi_\tau(F_{\text{OLS},h}) = \int \frac{\tau - \mathbb{1}\{y \le \psi_\tau(F)\}}{f_Y(\psi_{\tau,F})} \frac{x - \chi(F)}{\text{Var}_F[X]}\, \mathrm{d}F(x,y), \tag{1.19}$$

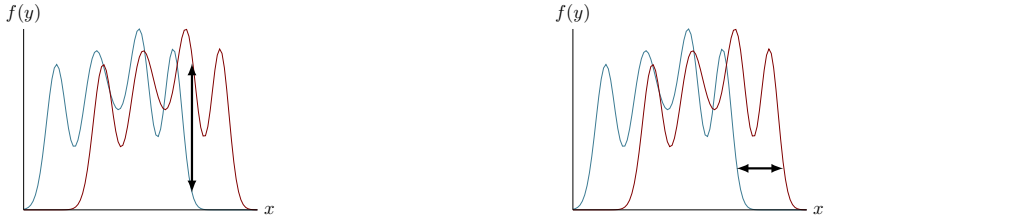
which is clearly different from α defined in eq. (1.18). That is, the infinitesimal changes in the quantile ψ_τ of Y along two paths of counterfactual distributions $F_{\text{OLS},h}$ and $F_{\text{LSH},h}$ are different.

Likewise, the local effect of changing the mean χ of X along the scores of the location shift model $F_{\text{LSH},h}$ on the mean μ of Y, is also given by the covariance of the influence function $\widetilde{\mu}_F$ and the score $\frac{d}{dh}\log f_{\text{LSH},h}$ of the model. From the results of [30] (or integration by parts) it follows that

$$\tfrac{d}{dh}\mu(F_{\text{LSH},h}) = \int \left[\tfrac{d}{dx} \int [y - \mu(F)]\, \mathrm{d}F_{Y|X}(y|x) \right] \mathrm{d}F_X(x), \tag{1.20}$$

which is clearly different from the OLS coefficient β in eq. (1.11), proving our claim. In other words, the curves of counterfactual distributions $F_{\text{OLS},h}$ and $F_{\text{LSH},h}$, that both contain the status quo distribution F_0 and are both parametrized by the change h in the mean χ of X from the status quo value $\chi(F_0)$, are different in general. Somewhat surprisingly, these curves coincide when the initial condition $F_{X,0}$ is a Gaussian.

In this paper we show that counterfactual distributions $F_{\text{OLS},h}$ and $F_{\text{LSH},h}$ are both *gradient flow curves* of the mean functional $\chi(F)$, but with respect to different metrics of distance on $\mathcal{F}$. The Wasserstein metric forces local changes in the distribution to be a continuous transportation of mass within the sample space $\mathcal{X}$. Wasserstein gradient flows minimize the distance in $\mathcal{X}$ that mass must be transported over. By contrast, the information metric is ignorant of the topology of $\mathcal{X}$ and allows mass to be moved discontinuously in $\mathcal{X}$. Information gradient flows minimize the total amount of mass that must be created and destroyed in $\mathcal{X}$.



(a) Fisher-Rao information distance $\mathfrak{g}_{\text{F}}$ (b) Wasserstein transport distance $\mathfrak{g}_{\text{K}}$

Figure 2: Vertical and horizontal metrics between probability measures

The location shift curve $F_{\text{LSH},h}$ in (1.17) turns out to be the most rapid way of changing the mean χ with respect to the Wasserstein distance between probability distributions. The exponential tilting curve $F_{\text{OLS},h}$ in (1.15) turns out to be the most rapid way of changing the mean χ with respect to the Fisher-Rao information distance between probability distributions.

In Section 2 we will explain the mathematical relationship between the gradient scores of the information and Wasserstein metrics and show how to obtain *transportation* curves of counterfactual distributions that generalize (1.17) by considering infinitesimal gradient changes in an arbitrary scalar functional ν of F.

1.4.3 **Instrumental variables regression (IV)** We now discuss the counterfactual distributions of the instrumental variables estimator. IV regression is an econometric technique used

in the context of the linear structural model

$$Y = \beta_{\text{IV},0} + \beta_{\text{IV}} X + \epsilon, \quad \text{Cov}\,(Z, \epsilon) = 0 \tag{1.21}$$

to find the effect of changes in the distribution of an endogenous covariate X on the outcome variable Y. The idea is to induce exogenous variation in the distribution of X with an instrumental variable Z in order to bypass the endogeneity in the process that generates the joint distribution F_{XY}. Endogeneity in F_{XY} means that the OLS coefficient $\beta(F_{XY})$ is not necessarily informative for the causal relationship between X on Y of interest to the researcher. The IV coefficient $\beta_{\text{IV}}(F_{XYZ})$ is commonly used in empirical work as an alternative to β, and is commonly interpreted the same way as the OLS coefficient β but with the endogeneity effects excluded.

Let $\zeta(F) = \int z \,\mathrm{d}F$ and $\widetilde{\zeta}_F(z) = z - \zeta(F)$ denote the mean of the instrument as a functional of the state of nature F_{XYZ} and its influence function. The nonparametric (nonlinear and nonstructural) counterfactual interpretation of the IV coefficient

$$\beta_{\text{IV}}(F) = \frac{\text{Cov}_F(Y,Z)}{\text{Cov}_F(X,Z)} = \int \widetilde{\mu}_F \frac{\widetilde{\zeta}_F}{\text{Var}_F[Z]}\,\mathrm{d}F \Big/ \int \widetilde{\chi}_F \frac{\widetilde{\zeta}_F}{\text{Var}_F[Z]}\,\mathrm{d}F = \int \widetilde{\mu}_F \frac{\widetilde{\zeta}_F}{\text{Cov}_F[X,Z]}\,\mathrm{d}F$$

follows from the von Mises formula (1.7) applied with the score function $v_{\text{IV},F} = \widetilde{\zeta}_F / \text{Cov}_F(X,Z)$, which is the normalized influence functions of the mean ζ of instrument Z.

By Proposition 2 applied with the score field $\{\widetilde{\zeta}_F\}_{F\in\mathcal{F}}$, there exists a model of counterfactual distributions $F_{\text{IV},t}$ containing the status quo distribution F_0 and parametrized by the change t in the mean ζ of Z, such that

$$\beta_{\text{IV}}(F_{\text{IV},t}) = \frac{d\mu}{dt} \Big/ \frac{d\chi}{dt}(F_{\text{IV},t}) = \frac{d\mu}{d\chi}(F_{\text{IV},t}) \tag{1.22}$$

everywhere on this curve. If we reparametrize this model by the change $h(t) = \chi(F_{\text{IV},t}) - \chi(F_0)$ in the mean χ of X (as in the counterfactual distributions $F_{\text{OLS},h}$ and $F_{\text{LSH},h}$ above), we then have $\frac{d}{dh}\mu(F_{\text{IV},h}) = \beta_{\text{IV}}(F_{\text{IV},h})$ for all h by the implicit function theorem. In other words, β_{IV} measures the effect on the mean μ of Y of the change the mean χ of X along the path of counterfactual distributions $F_{\text{IV},h}$. The counterfactual distributions $F_{\text{IV},h}$ are obtained by changing the marginal distribution F_Z along the scores $\widetilde{\zeta}_F$ while holding the conditional distributions $f_{XY|Z} = f_{Y|XZ} f_{X|Z}$ fixed, thereby changing the joint marginal distribution F_{XY}.

The effect of manipulating the mean of X with an instrument Z on the τ-quantile of Y is given by the covariance of the influence function $\widetilde{\psi}_{\tau,F}$ of the quantile functional and the score function $v_{\text{IV},F}$ of the path of counterfactual distributions $F_{\text{IV},h}$:

$$\tfrac{d}{dh}|_{h=0}\psi_\tau(F_{\text{IV},h}) = \int \frac{\tau - \mathbb{1}\{y \le \psi_\tau(F_0)\}}{f_Y(\psi_{\tau,F_0})} \frac{z - \zeta(F_0)}{\text{Cov}_{F_0}(X,Z)}\,\mathrm{d}F_0.$$

In the completely nonparametric model $\mathcal{F}$, the counterfactual distributions $F_{\text{IV},h}$ violate the Oaxaca-Blinder condition because the outcome assignment $F_{Y|X}$ may change with h. The IV perturbation score can be decomposed into $\widetilde{\zeta}_F = \text{E}_F[\widetilde{\zeta}_F(Z)|X] + \{\widetilde{\zeta}_F - \text{E}_F[\widetilde{\zeta}_F(Z)|X]\}$, where only the first term creates variation in the marginal distribution of X, and the second term can be understood to alter the outcome assignment $F_{Y|X}$. Therefore, we may define a modified IV coefficient with the OB property as:

$$\beta_{\text{IVX}} \coloneqq \frac{\text{Cov}_F(Y, \text{E}_F[Z|X])}{\text{Cov}_F(X,Z)} = \int \widetilde{\mu}_F(y) \left[\text{E}_F[\widetilde{\zeta}_F(Z)|X = x] / \text{Cov}_F(X,Z)\right] \mathrm{d}F.$$

The counterfactual distributions generated by the scores $v_{\text{IVX},F} = \text{E}_F[\widetilde{\zeta}_F(Z)|X] / \text{Cov}_F(X,Z)$ produce the same exogenous variation in the marginal distribution F_X while holding the outcome assignment $F_{Y|X}$ fixed.

To conclude, by revisiting empirical interpretation of common estimators with von Mises calculus of the estimand statistical functionals, we find that OLS and IV estimates have a nonparametric counterfactual interpretation as the effects of small changes in the mean of a covariate X on the mean of an outcome Y, but differ in the counterfactual distributions of these changes. Neither OLS nor IV counterfactual distributions are location shifts, but are exponential tilts of the marginal status quo distributions of X and Z respectively. Moreover, instruments with different joint distributions F_{XYZ} produce different marginal counterfactual distributions $F_{XY,h}$ and therefore different scalar effects $\Delta_h\psi(F_Y)$.

1.5 Effect of changing moments and descriptive statistics; GMM. The original motivating example for this paper is from Gentzkow and Shapiro (2015) [36] and Andrews, Gentzkow and Shapiro (AGS 2017, 2018) [6, 7], who consider functionals defined with complex structural models[9]. The problem AGS are concerned with is that the estimates in these models have a nontransparent relationship to the distribution of the data. Toward solving this problem, AGS propose formal measures of the relationship between structural parameter estimates and general descriptive statistics of the data.

The setup of AGS is a GMM vector estimand $\theta(P)$ that minimizes the population criterion function

$$\theta_W(P) \coloneqq \arg\min_\theta g_\theta(P)' W_P g_\theta(P), \quad \text{where} \quad g_\theta(P) \coloneqq P[g_\theta] \coloneqq \int_{\mathcal{X}} g_\theta(x)\,\mathrm{d}P(x) \tag{1.23}$$

is a vector of moments of $g : \Theta \times \mathcal{X} \to \mathbb{R}^r$ with respect to the distribution P that also depend on a vector of parameters $\theta \in \Theta \subset \mathbb{R}^p$, and W_P is a positive definite weighting matrix that depends only on P. The minimization problem (1.23) implicitly defines the vector of GMM functionals $\theta_W : \mathcal{P} \to \mathbb{R}^p$, whose components can be taken as the outcome functional ψ in the counterfactual analysis studied in the present paper. The core assumption in GMM estimation is that there exists a parameter value $\theta(P)$ that sets all r moment conditions to zero:

$$g_{\theta(P)}(P) = 0 \qquad \text{for all} \qquad P \in \mathcal{P}. \tag{1.24}$$

When the number of parameters $p < r$ is smaller than the number of moment conditions, (1.24) is a restriction on the data distributions $P \in \mathcal{P}$. On the restricted model, all functionals $\theta_W(P)$ coincide for all choices of the weighting functionals W, so can be denoted simply by $\theta(P)$. On the unrestricted model, different weighting matrices in (1.23) define different GMM functionals $\theta_W(P)$ because the criterion functions can have different minima and critical values. Because structural economic models never describe the real world exactly, the latter situation is much more likely to be the case for a given dataset.

The problem is to understand how $\theta_W(P)$ depends on descriptive statistics of the data distribution P. AGS define the sensitivity of the estimand $\theta(P)$, at a status quo distribution P_0 that is assumed to satisfy the moment conditions (1.24), to the fixed vector of moments $g_{|\theta=\theta_0}(P)$ corresponding to the status quo value of the parameter $\theta_0 = \theta(P_0)$, as:

$$\Lambda = -(G'WG)^{-1}G'W, \tag{1.25}$$

where G is the Jacobian matrix of the moment vector $\theta \mapsto g_\theta(P)$ whose θ-argument is evaluated at the status quo value θ_0, and the distributional P-argument of both W and G is evaluated at the status quo distribution P_0.

AGS [6] provide the following interpretation of the quantity Λ:

> "Intuitively, Λ is a local approximation to the mapping from moments $[g_{|\theta=\theta_0}(P)]$ to estimated parameters $[\theta(P_0)]$. A reader ... can use Λ to predict ... [changes in $\theta(P)$], provided

[9] To make the distinction between the structural estimation setting of AGS and the Oaxaca-Blinder analysis setting, we use notation $P \in \mathcal{P}$ for the joint probability distribution of the data X.

she can form a guess ... [of the changes in the moments $g_{|\theta=\theta_0}(P)$]".

The objective of [6] is to understand how $\theta(P_0)$ changes in response to a variation in a descriptive statistic $\nu : \mathcal{P} \to \mathbb{R}$ of the data distribution. In the present paper, we call this a counterfactual effect $\Delta_{\nu,h}\theta$ of a change h in ν on θ. As we showed in the case of the mean χ of X, there are many such sensitivities (i.e. infinitesimal counterfactual effects), enumerated by different scores v_P in von Mises formula (1.7).

The result obtained by AGS is the chain rule of the pathwise derivative (1.7), applied with the influence function $\widetilde{\theta}_P(x) = \Lambda_P g_{\theta(P)}(x)$ of the GMM functional $\theta(P)$ and the influence function $\widetilde{g}_{|\theta=\theta_0,P}(x) = g(\theta_0, x) - g_{\theta_0}(P)$ of the θ_0-moment functional $g_{|\theta=\theta_0}(P)$:

$$\tfrac{d}{d\nu}_{|P=P_0}\theta(P) = \int \widetilde{\theta}_{P_0}(x)\, v_{P_0}(x)\, \mathrm{d}P_0 = \int \Lambda_{P_0} g(\theta_0, x)\, v_{P_0}(x)\, \mathrm{d}P_0 = \Lambda_{P_0}\left[\tfrac{d}{d\nu} g_{|\theta=\theta_0}(P)\right]_{P=P_0},$$

which says that to find the local effect $\frac{d}{d\nu}\theta(P)$ on the GMM parameter, it is sufficient to find the effect on the (state of nature dependent) vector of moments $g_{|\theta=\theta_0}(P)$. In this paper, we provide the tools to study the infinitesimal effects $\frac{d}{d\nu}\theta(P)$ and $\frac{d}{d\nu} g_{|\theta=\theta_0}(P)$ of a change in a functional ν along different scores and the corresponding noninfinitesimal effects and distributions.

This paper extends AGS idea along several dimensions. We show that on the unrestricted model $\mathcal{P}$, the infinitesimal change $\frac{d}{d\nu}\theta_W(P)$ depends not only on the perturbation of the moments $\frac{d}{d\nu} g_{|\theta=\theta_0}(P)$, but also on the local changes in the second derivative $\partial^2_\theta g(\theta, P)$ of the moments and in the weighting functional W_P. Moreover, local changes in the distribution P and in the parameter $\theta_W(P)$ can be connected and integrated to obtain global changes in the distribution P_h and in the structural parameter $\Delta_{\nu,h}\theta_W$.

Proposition 3. On the unrestricted model $\mathcal{P}$, the influence function of the GMM functional θ_W defined in eq. (1.23) is

$$\begin{aligned}
&\widetilde{\theta}_{W,P}(x) = \\
&- \left[\left(P[g(\theta_{W,P})]^T W_P \otimes I_p\right) P\left[\partial_\theta \operatorname{vec}([\partial_\theta g(\theta_{W,P})]^T)\right] + P\left[\partial_\theta g(\theta_{W,P})\right]^T W_P\, P\left[\partial_\theta g(\theta_{W,P})\right] \right]^{-1} \\
&\quad \times \left\{ \left(P\left[g(\theta_{W,P})\right]^T W_P \otimes I_p\right) \left[\operatorname{vec}\left([\partial_\theta g(\theta_{W,P}, x)]^T\right) - P\left[\operatorname{vec}\left([\partial_\theta g(\theta_{W,P})]^T\right)\right] \right] \right. \qquad (1.26) \\
&\quad \left. + P\left[\partial_\theta g(\theta_{W,P})\right]^T \widetilde{W}_P(x)\, P\left[g(\theta_{W,P})\right] + P\left[\partial_\theta g(\theta_{W,P})\right]^T W_P \left(g(\theta_{W,P}, x) - P\left[g(\theta_{W,P})\right]\right) \right\}
\end{aligned}$$

where $\widetilde{W}_P$ is the influence function of the weighting functional W_P.

Suppose a scalar parameter ν defined on $\mathcal{P}$ can be changed locally at each counterfactual distribution $P \in \mathcal{P}$ along the score v_P. Suppose the collection of scores $\{v_P\}_{P\in\mathcal{P}}$ satisfies a local Lipschitz condition. Then at any status quo point $P_0 \in \mathcal{P}$, there exists a unique path of counterfactual distributions $\{P_h\}_{h\in J}$ defined on an open interval J containing 0, such that $\nu(P_h) = \nu(P_0) + h$ is the parameter of the path, and its score is proportional to v_{P_h} for every $h \in J$. Moreover, the change in the value of the GMM parameter $\theta_W(P)$ along this path of counterfactual probability distributions P_h is

$$\Delta_{\nu,h}\theta_W(P_0) = \theta_W(P_h) - \theta_W(P_0) = \int_0^h \int_{\mathcal{X}} \widetilde{\theta}_{W,P_t}\left[v_{P_t}/\operatorname{Cov}_{P_t}[\widetilde{\nu}_{P_t}, v_{P_t}]\right] \mathrm{d}P_t\, \mathrm{d}t,$$

where $\widetilde{\theta}_{W,P}$ is the influence function of the GMM functional given above, and $\widetilde{\nu}_P$ is the influence function of the descriptive statistic ν.

The influence function of a statistical parameter is closely related to the asymptotic distribution of regular estimators of that parameter. The general asymptotic distribution of the overidentified GMM estimator is discussed implicitly in Imbens (1997) [46], and derived explic-

itly in Hall and Inoue [39]. Here we derive the general form of the influence function.

The main limitation of the sensitivity measure Λ for structural parameters defined by AGS in [6] and the global effects $\Delta_{\nu,h}\theta_W$ defined in Proposition 3 is that the scores v_P for perturbing the value of the descriptive statistic ν must be provided by the reader. The influence function $v_P = \widetilde{\nu}_P$ is one natural choice of these scores, based on all previous examples of this paper. In Section 2 we derive the scores v_P featured in Proposition 3 from the descriptive statistic ν via local optimality conditions with respect to different criteria of optimality. This provides another extension to the idea of AGS of how one can relate changes in a descriptive statistic to changes in a structural parameter.

2. Score counterfactuals

As we discussed in the Introduction and examples of Section 1, in a number of empirical papers in economics, one finds explicit or implicit discussions of the effect of changes in one scalar statistical functional ν on another statistical functional ψ. Formally this means that one is interested in the effects on ψ of the changes in the status quo probability distribution, which are somehow specified by the counterfactual values of the parameter ν. The situation is analogous to, in fact some of the obtained results are corollaries of, the idea of *the hardest submodel* formulated in Stein's (1956) paper [66] and developed rigorously by Levit (1974, 1975, 1976) in [50, 51, 47] and many others. We adapt Stein's idea to counterfactual analysis as follows: *Given a small increment h in the status quo value of the policy functional $\nu(P_0)$, it is required to find the distribution $P_{\nu,h}$ that "deviates minimally" from the status quo state of nature P_0 and has $\nu(P_{\nu,h}) = \nu(P_0) + h$.* For the purposes of economic counterfactual analysis, we interpret "minimally" as "in the cheapest fashion" so that the "hardest" submodel becomes the "optimal" one for e.g. policy analysis purposes.

Suppose $\mathcal{P}$ is a collection of counterfactual probability distributions for the random variable X on the sample space $(\mathcal{X}, \mathcal{A})$, and let $P_0 \in \mathcal{P}$ denote the status quo distribution (e.g. the true state of nature). Let us setup a basic geometric structure on $\mathcal{P}$ (i.e. a manifold[10]) in the spirit of semiparametric efficiency theory [62, 53, 66, 29, 47, 10, 72, 13, 71] and information [43, 52, 2, 9, 3, 61, 16, 19, 57, 37] and optimal transportation [35, 12, 5, 4] geometries. The main economic idea here is to allow a general metric of "cost" of counterfactual states of nature $P_h \in \mathcal{P}$. The main technical idea is to work around the difficulties imposed by nonlinearities in the model $\mathcal{P}$ and its functionals $\nu, \psi : \mathcal{P} \to \mathbb{R}$ by working in an infinitesimal neighborhood of a point P. Locally, the model $\mathcal{P}$ and the functionals ν, ψ are *linear* making infinitesimal counterfactual analysis tractable. Furthermore, to construct global counterfactual distributions and their parameters we can integrate consecutive local changes in the probability distribution and its functionals. In this section we consider the local analysis, in Section 3 we consider the global analysis.

Tangent space At each point $P \in \mathcal{P}$, we consider a collection of smooth curves $\{P_t\}_{0<t<\epsilon} \subset \mathcal{P}$ that pass through P at $t = 0$ and possess a score function v_P at P. This means:

$$\int_{\mathcal{X}} \left[t^{-1}(\sqrt{dP_t} - \sqrt{dP}) - \tfrac{1}{2} v_P \sqrt{dP}) \right]^2 \to 0 \quad \text{as } t \to 0, \tag{2.1}$$

where dP_t, dP are densities with respect to some dominating measure (van der Vaart's notation). In other words, the score v_P is the velocity of the curve P_t in the embedding of the model $\mathcal{P}$ into the space H_2 of square roots of measures [see 54, p. 112]. The score function $v_P : \mathcal{X} \to \mathbb{R}$ of a

[10] To focus on the much higher level topic of empirical counterfactual analysis of this paper, we introduce very briefly only the necessary geometric notation and completely ignore the fundamental notion of charts. See e.g. Carmo (1976, 1992) [17, 18] and Lang (1999) [48] for complete and rigorous expositions of definitions of geometry. See Pistone and Sempi (1995) [61], Cena and Pistone (2007) [19], Fukumizu (2009) [34], Grasselli (2010) [37] and Newton (2012) [57] for rigorous constructions of nonparametric statistical manifolds.

curve P_t is typically computed by differentiating the log density as in eq. (1.8), and has finite second moment and zero expectation under P [see e.g. 47, 13, 71].

The *tangent space* $\mathcal{T}_P$ to the model $\mathcal{P}$ at the point P is a linear subspace of $L_0^2(P)$ composed of *scores* v_P of all regularly parametrized submodels $\{P_t\}_{0<t<\epsilon} \subset \mathcal{P}$ through the point P.

Metric A *metric* $\mathfrak{g}$ is a collection of complete inner products $\{\mathfrak{g}_P(\cdot,\cdot)\}_{P\in\mathcal{P}}$ on the tangent spaces $\mathcal{T}_P$ of the model $\mathcal{P}$. A metric induces Hilbert norms $\|v_P\|_{\mathfrak{g},P}^2 = \mathfrak{g}_P(v_P, v_P)$ on score functions $v_P \in \mathcal{T}_P$, and the geodesic distance function

$$\mathfrak{g}(P_h, P_0) \coloneqq \min_{t\mapsto P_t} \int_0^h \|v_{P_t}\|_{\mathfrak{g},P_t}\, \mathrm{d}t \tag{2.2}$$

on the set of counterfactual distributions $\mathcal{P}$. The metric $\mathfrak{g}_P(\cdot,\cdot)$ can be thought of as the infinitesimal representation of the distance function $\mathfrak{g}(\cdot,\cdot)$. Not every distance on the space of probability measures has the infinitesimal representation (2.2). The infinitesimal structure of the distance on $\mathcal{P}$ is central to the analysis of this paper, we consider only metric distances.[11]

Pathwise derivative The *pathwise derivative* of a functional $\nu : \mathcal{P} \to \mathbb{R}$ at a point P is a continuous linear map $\mathrm{d}\nu_P : \mathcal{T}_P \to \mathbb{R}$ that satisfies

$$\mathrm{d}\nu_P(v_P) = \tfrac{d}{dt}_{|t=0}\nu(P_t) = \lim_{t\to 0} t^{-1}\big[\nu(dP + tv_P dP + o(t)) - \nu(P)\big] \tag{2.3}$$

for every regular submodel P_t with the score v_P, and every score $v_P \in \mathcal{T}_P$. It is important for the analysis of this paper that a functional admits a local linear approximation. Luckily all functionals that can be estimated at the parametric $\sqrt{n}$-rate with some uniformity in the asymptotic distributions are pathwise differentiable [72].

Influence function, gradient The *influence function* $\widetilde{\nu}_P$ of a functional ν at a point P is the Riesz representation score of its pathwise derivative $\mathrm{d}\nu_P$ with respect to the $L^2(P)$ inner product on $\mathcal{T}_P$. We refer to Dudley (2002) [28, p174] for Hilbert space theory.

The $\mathfrak{g}$-*gradient* of ν, denoted by $\nabla_{\mathfrak{g}}\nu_P$, is the Riesz representation score of its pathwise derivative with respect to the inner product $\mathfrak{g}_P$ of the metric on $\mathcal{T}_P$. This means that the pathwise derivative must satisfy

$$\tfrac{d}{dt}_{|t=0}\nu(P_t) = \mathfrak{g}_P\big(\nabla_{\mathfrak{g}}\nu_P, v_P\big) \tag{2.4}$$

for every regular path P_t with score v_P, and every score $v_P \in \mathcal{T}_P$.

In what follows, we will assume that all paths P_t, functionals ν, ψ and metrics $\mathfrak{g}$ are smooth.

The main idea of this paper is to use the Riesz representation theorem to regularize the Problem 1 of specifying counterfactual distributions P_{ν,h,P_0} with changes in a scalar functional $\nu(P_0)$ of the probability distribution. Riesz representation theorem says that for a given notion of local (and global) distance $\mathfrak{g}$ on $\mathcal{P}$, there is a unique natural score function $\nabla_{\mathfrak{g}}\nu_P \in \mathcal{T}_P$ at each point P to associate with the functional. The following result provides a useful characterization of the Riesz representation score:

Proposition 4 (Optimality of gradient scores)**.** Given a pathwise differentiable functional ν and a metric of distance $\mathfrak{g}$ defined on the set of counterfactual distributions $\mathcal{P}$ with tangent spaces $\mathcal{T}_P$, the normalized gradients

$$\nabla_{\mathfrak{g}}^1\nu_P(x) \coloneqq \nabla_{\mathfrak{g}}\nu_P(x)/\mathrm{Cov}_P\big[\widetilde{\nu}_P(X), \nabla_{\mathfrak{g}}\nu_P(X)\big] \tag{2.5}$$

[11] Mathematical literature has extended the notion of gradient flows to the much more general setting of metric spaces. See Ambrosio, Gigli and Savare (2008) [4]. We do not consider such extensions in this paper.

uniquely satisfy the local optimality conditions

$$\min_{v_P \in \mathcal{T}_P} \|v_P\|_{\mathfrak{g},P} \quad \text{s.t.} \quad \lim_{t\to 0} t^{-1}\big[\nu(dP + tv_P dP + o(h)) - \nu(P)\big] = 1$$

of increasing the value of the functional ν, while minimizing the cost of the deviation measured by the metric $\mathfrak{g}$, among all local perturbations v_P with the same effect on ν.

Proposition 4 can be interpreted as follows. If we measure the "cost" of a counterfactual distribution P_h by its distance $\mathfrak{g}(P_h, P_0)$ from the status quo distribution P_0, and measure the "utility" of the counterfactual distribution by the change in the functional $\nu(P_h) - \nu(P_0)$, then Proposition 4 says that the optimal counterfactual distributions in a neighborhood of P_0 are those in the direction of the gradient $\nabla_{\mathfrak{g}}\nu_{P_0}$. One way to interpret the gradient $\nabla^1_{\mathfrak{g}}\nu_P$, is to think of it as the local compensated demand in $\mathcal{P}$ for the policymaker with utility ν and costs $\mathfrak{g}$. Normalization by the covariance term in eq. (2.5) and eq. (1.9) is the result of looking at infinitesimal changes in ν of a unit magnitude (reflected with the superscript ∇^1 in gradient score notation). Proposition 4 provides a score function at every point $P \in \mathcal{P}$, and together with Proposition 2 allow counterfactual analysis to be conducted with a scalar policy functional ν, as frequently done informally in empirical papers, by choosing a metric:

Corollary 5 (Gradient flow curve of counterfactual distributions)**.** Suppose ν is pathwise differentiable functional and that $\mathfrak{g}$ is a metric on $\mathcal{P}$. Provided that the gradient score field $\{\nabla_{\mathfrak{g}}\nu_P\}_{P\in\mathcal{P}}$ satisfies a local Lipschitz condition, at any status quo distribution $P_0 \in \mathcal{P}$, there exists a unique regular parametric model $\{P_{\mathfrak{g},\nu,h,P_0}\}_{h\in J}$ defined on an open interval J containing 0, parametrized by the change $h = \nu(P_h) - \nu(P_0)$ in ν, and with the scores $\frac{d}{dh}\log dP_h = \nabla^1_{\mathfrak{g}}\nu_{P_h}$ for every $h \in J$.

Next we describe the gradient scores of two important distance functions on the space of probability measures: the Fisher-Rao information metric that arises naturally in asymptotic statistics, and the 2-Wasserstein metric of the Monge-Kantorovich theory of optimal transportation of mass. These metrics have been at work implicitly in the examples from empirical literature discussed in Section 1.

2.1 Counterfactuals via Fisher-Rao information metric gradient flow. The information metric, denoted by $\mathfrak{g}_{\text{F}}$, is just the collection of $L^2(P)$ inner products

$$\mathfrak{g}_{\text{F},P}(v_P, w_P) = \int_{\mathcal{X}} v_P\, w_P \, \mathrm{d}P = \mathrm{Cov}_P[v_P, w_P], \qquad v_P, w_P \in \mathcal{T}_P,\ P \in \mathcal{P} \tag{2.6}$$

on the tangent spaces $v_P, w_P \in \mathcal{T}_P$ of $\mathcal{P}$. The corresponding global distance function on $\mathcal{P}$ is

$$\mathfrak{g}_{\text{F}}(P_0, P_h) = 2\arccos\Big(1 - 1/2\int_{\mathcal{X}}(\sqrt{dP_0} - \sqrt{dP_h})^2\Big), \qquad P_0, P_h \in \mathcal{P} \tag{2.7}$$

and is basically the L^2 distance between the square roots of the density functions of the probability measures. In this sense, the information metric is said to measure the "vertical" distance between probability distributions and is ignorant of the topology of the sample space $(\mathcal{X}, \mathcal{A})$.

The *information gradient* $\nabla_{\text{F}}\nu_P$ of a functional ν is its influence functions $\widetilde{\nu}_P$. This gradient shows the optimal variation in the density of the distribution $\dot{f}_P = \widetilde{\nu}_P f_P$ that increases the value of the functional, while minimizing the statistical dissimilarity of counterfactual distributions with P [47]. The $\mathfrak{g}_{\text{F}}$-gradient flow curve $P_{\text{F},\nu,h}$ can be considered the canonical *hardest submodel* for estimating parameter ν, because all of its scores are information gradients of ν and it is unique under regularity conditions. The hardest submodel $P_{\text{F},\nu,h}$ contains the least statistical information for estimating the parameter ν and inflicts the greatest dispersion in the asymptotic distribution of its regular estimators $\widehat{\nu}$, known as the *efficiency bound*. This property links the

asymptotic distribution of regular estimators at P_0 on the entire model $\mathcal{P}$ with the counterfactual distributions along the information gradient flow curve P_{F,ν,h,P_0}.

The following result was stated informally in Gentzkow and Shapiro (2015) [36], and is a corollary of Stein (1956) [66] and many subsequent papers [e.g. 47, 10, 72].

Theorem 6 (Information gradient flow effects via asymptotic distribution of regular estimators)**.** Assume that model $\mathcal{P}$ allows arbitrary misspecification and has nonparametric tangent sets $\mathcal{T}_P = L_0^2(P)$. Let data $X_1, \ldots, X_n$ be a random sample from the distribution $P \in \mathcal{P}$. Suppose estimator sequences $\widehat{\psi}_n = \widehat{\psi}_n(X_1, ..., X_n)$ and $\widehat{\nu}_n = \widehat{\nu}_n(X_1, ..., X_n)$ have a linear asymptotic representation:

$$\sqrt{n}\begin{bmatrix}\widehat{\psi}_n - \psi(P) \\ \widehat{\nu}_n - \nu(P)\end{bmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix}\widetilde{\psi}_P(X_i) \\ \widetilde{\nu}_P(X_i)\end{bmatrix} + o_{P^n,n}(1), \quad \widetilde{\psi}_P, \widetilde{\nu}_P \in L_0^2(P). \tag{2.8}$$

Define the functionals $\psi(P), \nu(P)$ as the large sample limits of the estimators on $\mathcal{P}$. Furthermore, suppose that for every regular parametric path P_h satisfying (2.1):

(i) the remainder terms $o_{P^n(h),n}(1)$ converge to zero uniformly in h as $n \to \infty$;

(ii) the norms $\int \widetilde{\psi}^2_{P(h)} dP_h$ and $\int \widetilde{\nu}^2_{P(h)} dP_h$ are continuous in h;

(iii) the score field $\{\widetilde{\nu}_P\}_{P\in\mathcal{P}}$ of influence functions of the estimator satisfies a local Lipschitz condition.

Then (i) The functionals ψ, ν are pathwise differentiable, and their information gradients $\nabla_{\text{F}}\psi_P, \nabla_{\text{F}}\nu_P$ are equal to the influence functions $\widetilde{\psi}_P, \widetilde{\nu}_P$ of the estimators. (ii) For any status quo distribution $P_0 \in \mathcal{P}$, there exists a unique regular parametric model $\{P_{\text{F},\nu,h,P_0}\}_{h\in J}$ defined on an open interval J containing 0, parametrized by the change $h = \nu(P_{\text{F},\nu,h,P_0}) - \nu(P_0)$ in ν, and with the scores $\frac{d}{dh}\log dP_{\text{F},\nu,h,P_0} = \widetilde{\nu}_P/\text{Var}_P[\widetilde{\nu}_P(X)]$ for every $h \in J$. (iii) The local effects of changing ν along the hardest submodel P_{F,ν,h,P_0} on ψ are given by

$$\tfrac{d}{dh}\psi(P_{\text{F},\nu,h,P_0}) = \text{Cov}_{P_h}\big[\widetilde{\psi}_{P_h}(X), \widetilde{\nu}_{P_h}(X)\big]/\text{Var}_{P_h}\big[\widetilde{\nu}_{P_h}(X)\big], \tag{2.9}$$

which is the covariance of the asymptotic distribution at P_h of the joint estimator $(\widehat{\psi}_n, \widehat{\nu}_n)$ normalized by the asymptotic variance of $\widehat{\nu}_n$. (iv) The global effects of changing ν along the hardest submodel P_{F,ν,h,P_0} on ψ are given by

$$\begin{aligned}\Delta_{\text{F},\nu,h,P_0}\psi(P_0) &= \psi\big(P_{\text{F},\nu,h,P_0}\big) - \psi(P_0) \\ &= \int_0^h \text{Cov}_{P_t}\big[\widetilde{\psi}_{P_t}(X), \widetilde{\nu}_{P_t}(X)\big]/\text{Var}_{P_t}\big[\widetilde{\nu}_{P_t}(X)\big]\ \text{d}t,\end{aligned} \tag{2.10}$$

which is a linear combination of the local effects.

Theorem 6 says that approximate counterfactual effects are available in every empirical paper that reports standard errors based on the usual $\sqrt{n}$-asymptotic approximations as a byproduct of calculating the asymptotic distribution. Of course, asymptotic inference implicitly relies on the information metric and provides counterfactual effects only along the hardest model through the true distribution of the data (provided that the researcher calculates the accurate nonparametric asymptotic distribution as we have done for the GMM estimator in eq. (1.26)). It would actually be more accurate to say that asymptotic standard errors are a byproduct of the information counterfactuals [47].

The idea of the hardest submodel was presented informally in [66]. This idea of making counterfactual statements based on the asymptotic distribution was presented informally in [36]. The contribution of the present paper is to supply the precise geometric interpretation and implementation of both ideas and to find the technical regularity conditions from the semiparemtric efficiency literature and the nonparametric information geometry literature. The main difficulty

is to formalize the ordinary differential equation on the model $\mathcal{P}$ of probability measures that is determined by the score field of influence functions. Theorem 6 generalized the OLS example of Section 1.4.1 which considered the means of an outcome Y and a covariate X, and Proposition 3 that considered GMM and descriptive statistics.

Example 2.1 (The effect of changing quantiles on moments)**.** Returning to examples of Sections 1.3 to 1.5 that described the scalar effects of changing the mean of a covariate along the hardest submodel, we can now find geometrically *comparable* counterfactual effects of a counterfactual change in *any* functional by Proposition 4. Consider the effect of changing the quantiles of a distribution. Let $\psi_\rho(P) \coloneqq \int_{\mathcal{X}} \rho(x)\,\mathrm{d}P(x)$ be a moment functional with a general moment function $\rho : \mathcal{X} \to \mathbb{R}$, and let $\nu_\tau(P) \coloneqq F^{-1}_{X_j,P}(\tau)$ denote the τ-quantile of the jth covariate X_j when the joint distribution of data is P. The moment functional has influence function

$$\psi_\rho(P) \coloneqq \int_{\mathcal{X}} \rho(x)\,\mathrm{d}P(x), \qquad \widetilde{\psi}_{\rho,P}(x) = \rho(x) - \psi_\rho(P) \tag{2.11}$$

and the influence function of the quantile can be found in eq. (1.6). By Theorem 6, the effect of changing the quantile ν_τ along its hardest submodel for estimating the quantile on the moment ψ_ρ is the normalized asymptotic covariance of a joint estimator:

$$\tfrac{d}{dh}\psi_\rho(P_{\mathrm{F},\nu_\tau,h}) = \tfrac{f_{X_j}(\nu_{\tau,P})}{\tau(1-\tau)}\mathrm{E}_P\Big[\big(\rho(X) - \psi_\rho(P)\big)\big(\tau - \mathbb{1}\{X_j \le \nu_\tau(P)\}\big)\Big]. \tag{2.12}$$

Note that the effect scales linearly with the marginal density $f_{X_j}(\nu_{\tau,P})$ at the quantile. We shed light on this curious phenomenon of the asymptotic behavior of quantile estimates with the transportation gradient of the quantile functional in the next subsection.

To apply Proposition 4 and Corollary 5 to counterfactual analysis in practice, the researcher chooses a functional and a metric that best describe the hypothetical state of nature under examination and derives the gradient score. Since the influence functions of many parameters are known, and results that describe how to derive it are available [e.g. 45, and Theorem 8 below] this task typically amounts to finding a formula that maps the influence function into the gradient. The next result shows the effect of reweighting the information metric by a "cost distribution" term $\frac{dQ}{dP}$, so that the metric becomes

$$\mathfrak{g}_{\mathrm{F}_Q,P}(v_P, w_P) = \int_{\mathcal{X}} v_P\, w_P\, \mathrm{d}Q,$$

on the direction of the gradient flow curve:

Lemma 7 (Weighted information gradient scores)**.** Suppose $\nu : \mathcal{P} \to \mathbb{R}$ is a pathwise differentiable functional with influence functions $\widetilde{\nu}_P \in L^2_0(P)$. Suppose the cost distributions satisfy $Q_P \ll P$ for all $P \in \mathcal{P}$ and have uniformly bounded derivatives $0 < m \le \frac{dQ_P}{dP}(x) \le M < +\infty$ on $\mathcal{X}$. Then

$$\nabla_{\mathfrak{g}}\nu_P(x) = \Big[\widetilde{\nu}_P(x) - \textstyle\int_{\mathcal{X}} \widetilde{\nu}_P\, \frac{dP}{dQ_P}\,\mathrm{d}P \Big/ \int_{\mathcal{X}} \frac{dP}{dQ_P}\,\mathrm{d}P\Big]\frac{dP}{dQ_P}(x) \tag{2.13}$$

are the gradient scores for the weighted information metric $\mathfrak{g}_{\mathrm{F}_Q,P}(v_P, w_P) = \int v_P w_P\,\mathrm{d}Q_P$.

The effect of introducing the reweighting term $\frac{dQ}{dP}$ in the information metric on the gradient score is twofold: (i) the magnitude of mass creation/destruction is scaled by the $\frac{dQ}{dP}^{-1}(x)$ through the sample space $x \in \mathcal{X}$; (ii) the resulting score function is recentered to have zero mean under P. The rescaling effect provides a justification to our informal interpretation of the metric as the "cost" of counterfactual states of nature. I.e. the higher the cost $dQ_P(x)$ charged by the metric $\mathfrak{g}_{\mathrm{F}_Q}$ for creating/destroying mass at $x \in \mathcal{X}$, the smaller the contribution of the perturbation $\nabla_{\mathfrak{g}}\nu_P(x)$ to the distribution P at x. The information metric $\mathfrak{g}_{\mathrm{F}}$ can be interpreted as having

the uniform cost for creation/destruction of mass throughout the sample space. The fact that a large class of changes in the metric does not radically change the profile of the gradient, and consequently the counterfactual effects on the outcome functionals, provides some justification to the practice of reporting sensitivities (2.9) (i.e. information gradient flow effects) for structural estimates suggested in [36, 6, 7].

The next result shows how to compute the influence function $\widetilde{\psi}_P(z)$ by calculating the derivative of the functional along a special regular path that approximates the perturbation toward the point mass δ_z. The formula extends the calculation of von Mises (1947) [53] and Hampel (1974) [40] that recovers the influence function $\widetilde{\psi}_P(z)$ of a functional ψ by computing its derivative along the path $P_{t,z} = (1-t)P+t\delta_z$. Unfortunately the path $P_{t,z}$ is not regular and does not have a score function at $t = 0$, as shown in Appendix 4.4. This means that the von Mises calculation applies only to functionals that have stronger smoothness than the standard notion of pathwise differentiability in semiparametric efficiency theory. Ichichura and Newey (2017) [45] mollify the point mass in the von Mises calculation using a kernel, and recover the influence function from Riesz representation similarly to the way the value of a density function at a point is estimated from data. The result below is a partial generalization of this idea. We consider only absolutely continuous distributions P but allow all pathwise differentiable functionals ψ. Specifically, no assumption about the regularity of the influence function $\widetilde{\psi}_P$ is made other than the default $L_0^2(P)$ integrability. The theorem covers all pathwise differentiable functionals and all possible influence functions $\widetilde{\psi}_P \in L_0^2(P)$, so is the most general possible, but requires that the probability measure P has a continuous Lebesgue density. The theorem removes the continuity assumption about the influence function $\widetilde{\psi}_P$ by applying Lebesgue differentiation and approximation to the identity ideas from analysis. We refer to Stein and Shakarchi (2009) [67] for details of measure theory. We remark that the assumptions about the distribution P are probably not necessary either.

Theorem 8 (General von Mises influence function formula via approximations to the identity)**.** Let $\psi : \mathcal{P} \to \mathbb{R}$ be a pathwise differentiable functional on the nonparametric model $\mathcal{P}$ with influence functions $\widetilde{\psi}_P \in L_0^2(P)$. Suppose P_0 is an absolutely continuous probability measure with respect to the Lebesgue measure on $\mathbb{R}^d$ with a continuous density function f_0. Suppose K is a bounded probability density function with support in the unit ball $\{|x| \leq 1\} \subset \mathbb{R}^d$, and let

$$K_\delta(x) \coloneqq \delta^{-d} K(\delta^{-1}x), \quad \delta > 0 \tag{2.14}$$

$$K_{f_0,\delta,z}(x) \coloneqq \Big[\int_{\{f_0>\delta\}} K_\delta(x)\, dx\Big]^{-1} \mathbb{1}_{\{f_0>\delta\}}(z-x) K_\delta(z-x). \tag{2.15}$$

For $t \in (0,1)$, small $\delta > 0$ and $z \in \{f_0(x) > 0\} \subset \mathcal{X}$ consider the deviations $P_{t,\delta,z}$:

$$f_{t,\delta,z}(x) \coloneqq (1-t)f_0(x) + tK_{f_0,\delta,z}(x) \tag{2.16}$$

from the probability distribution P_0 toward the point-mass at z, specified in terms of its density function f_0 and approximation to the identity K_δ. Then the following *influence function formula* holds:

$$\widetilde{\psi}_{P_0}(z) = \lim_{\delta \to 0} \tfrac{d}{dt}_{|t=0} \psi(P_{t,\delta,z}) \tag{2.17}$$

for P_0-almost every $z \in \mathbb{R}^d$.

Proof. We outline the main ideas of the proof and provide the details in Appendix 4.4. The score of the path $t \mapsto P_{t,\delta,z}$ is

$$v_{0,\delta,z}(x) = \tfrac{d}{dt}_{|t=0} \log\Big\{f_0(x) + t\big[K_\delta(z-x) - f_0(x)\big]\Big\} = K_\delta(z-x)/f_0(x) - 1.$$

By pathwise differentiability of ψ at P_0, the derivative of the functional along the curve $t \mapsto P_{t,\delta,z}$ at $t = 0$ is given by the Riesz representation of its derivative mapping $d\psi_{P_0}$, i.e. the covariance inner product of the influence function $\widetilde{\psi}_{P_0}$ and the score function $v_{0,\delta,z}$:

$$\begin{aligned}
\tfrac{d}{dt}_{|t=0}\psi(P_{t,\delta,z}) &= \mathrm{d}\psi_{P_0}(v_{0,\delta,z}) \\
&= \int_{\mathcal{X}} \widetilde{\psi}_{P_0}(x)\, v_{0,\delta,z}(x)\, dP_0 \\
&= \int_{\mathcal{X}} \widetilde{\psi}_{P_0}(x) \left[K_\delta(z-x)/f_0(x) - 1\right] dP_0 \\
&= \int_{\mathrm{supp} P_0} \widetilde{\psi}_{P_0}(x)\, K_\delta(z-x)\, dx \\
&= (\widetilde{\psi}_{P_0} * K_\delta)(z).
\end{aligned}$$

The properties assumed about the kernels K_δ insure that it is an approximation to the identity in the sense that approximates a point mass at 0 in the and consequently

$$(\widetilde{\psi}_{P_0} * K_\delta)(z) \to \widetilde{\psi}_{P_0}(z) \qquad \text{as } \delta \to 0$$

for P_0-almost every $z \in \mathbb{R}^d$. □

2.2 Counterfactuals via Wasserstein transportation metric gradient flow. In some applications it is desirable to consider counterfactual distributions that imply an unambiguous counterfactual effect for each agent $x \in \mathcal{X} \subset \mathbb{R}^d$, e.g. Stock (1989) [68] specifies the counterfactual distribution P_h with an explicit transformation $X \mapsto X^*$ of the covariates; Firpo, Fortin, and Lemieux (2009) [30], discussed in section 1.4, consider a location shift of the marginal distribution of a covariate, $X_j \mapsto X_j^* = X_j + h$, $X_{-j}^* = X_{-j}$.

Transportation counterfactual distributions can be specified with gradient flow changes of the policy functional ν in the L^2-Wasserstein distance on the space of counterfactual distributions with finite second moments $\mathscr{P}_2(\mathbb{R}^d)$:

$$W_2(P_0, P_h) \coloneqq \Big(\min_{\gamma \in \Gamma(P_0, P_h)} \int_{\mathcal{X}\times\mathcal{X}} |x - x^*|^2 \, \mathrm{d}\gamma(x, x^*) \Big)^{1/2}, \tag{2.18}$$

where a *transport plan* $\gamma \in \Gamma(P_0, P_h)$ is a coupling (i.e. joint distribution) of the status quo and the counterfactual distributions P_0 and P_h, whose conditional distributions $\gamma_{X^*|X=x}$ describe the assignment of counterfactuals x^* to agents with the status quo covariate $x \in \mathcal{X}$. When the status quo distribution $P_0 = \varrho_0 \mathscr{L}^d$ has a density ϱ_0 with respect to the Lebesgue measure $\mathscr{L}^d$ on $\mathbb{R}^d$, the optimal assignment in eq. (2.18) is actually given by a transport map $x \mapsto x^*$ that is almost surely the gradient $\nabla_x \varphi$ of a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$. We will assume that all distributions have smooth densities and enough regularity to justify the manipulations in this section and refer to [5, 4] for the technical details.

The collection of smooth transport curves $P_t = \varrho_t \mathscr{L}^d$ in the W_2-distance is characterized by the *continuity* equation:

$$\tfrac{d}{dt}\varrho_t + \mathrm{div}(\varrho_t \boldsymbol{v}_t) = 0 \tag{2.19}$$

which says that the infinitesimal change $\frac{d}{dt}\varrho_t(x)$ in the density of agents at location x must be exactly matched by the net infinitesimal transport (i.e. flux) of agents into location x. The velocity vector field $\boldsymbol{v}_t : \mathcal{X} \to \mathbb{R}^d$ of the curve P_t measures the instantaneous rate $\boldsymbol{v}_t(x)$ of transport of agents at location $x \in \mathcal{X}$ at time t, and the divergence term $\mathrm{div}(\varrho_t \boldsymbol{v}_t)$ computes the resulting net instantaneous outflow of agents from x [see e.g. 32, Sec 5.5]. Equating these two terms in (2.19) prevents creation/destruction of density that is typical of smooth curves in

the information metric, and forces all changes in the density to be the result of a continuous transport of mass in the sample space $\mathcal{X}$. In terms of the tangent space $\mathcal{T}_P$, the score of the transport curve $t \mapsto P_t$ is the derivative of the log likelihood

$$v_{P_t} = \tfrac{d}{dt} \log \varrho_t = \tfrac{d}{dt} \varrho_t / \varrho_t = -\mathrm{div}(\boldsymbol{v}_t \varrho_t)/\varrho_t, \tag{2.20}$$

and the 2-Wasserstein metric inner-product on $\mathcal{T}_P$ is given by

$$\mathfrak{g}_{\mathrm{K},P}(v_P, w_P) = \int_{\mathcal{X}} \langle \boldsymbol{v}(x), \boldsymbol{w}(x) \rangle_{\mathbb{R}^d} \, dP(x) \tag{2.21}$$

for the canonical choice of velocity vector fields $\boldsymbol{v}, \boldsymbol{w} \in \overline{\{\nabla_x \varphi \,;\, \varphi \in C_c^\infty(\mathcal{X})\}}^{L^2(P;\mathbb{R}^d)}$ in the continuity equation (2.19) that produce the scores v_P, w_P. The fact that the Wasserstein distance (2.18) has an infinitesimal structure of a metric inner product is a nontrivial result of Benamou and Brenier (2000) [12]. The next result relates the gradient scores of the Wasserstein metric to influence functions.

Proposition 9 (Wasserstein gradient formula)**.** Suppose $\nu : \mathcal{P} \to \mathbb{R}$ is a pathwise differentiable functional with influence function $\widetilde{\nu}_P : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ in $L_0^2(P)$ that has a distributional derivative $\nabla_x \widetilde{\nu}_P$ in $L^2(P; \mathbb{R}^d)$. Suppose P has compact support and a smooth density $\varrho = \frac{dP}{d\mathscr{L}^d} \in C_c^1(\mathbb{R}^d)$. Then the gradient score of ν with respect to the 2-Wasserstein metric is

$$\nabla_{\mathrm{K}} \nu_P = -\mathrm{div}(\varrho \, \nabla_x \widetilde{\nu}_P)/\varrho. \tag{2.22}$$

We illustrate the direction of the Wasserstein gradient score with the following examples:

Example 2.2 (Mean)**.** Recall that the moment functional $\nu_\rho(P) = \int_{\mathcal{X}} \rho(x) \, \mathrm{d}P(x)$ with a moment function $\rho : \mathcal{X} \to \mathbb{R}$ has the influence function $\widetilde{\nu}_{\rho,P}(x) = \rho(x) - \nu_\rho(P)$. Provided that the moment function ρ is sufficiently smooth and P is sufficiently regular, the velocity vector field of the gradient transport for ν_ρ is $\boldsymbol{v} = \nabla_x \rho \in L^2(P; \mathbb{R}^d)$. In particular, we can derive the score of the infinitesimal effect α, defined in eq. (1.18), of shifting the marginal distribution of a scalar X and holding the conditional distribution of $Y|X$ fixed as in the Oaxaca-Blinder applications, by taking $\rho(x) = x$ so that the functional is the mean ν_1 of X as in Section 1.4. It follows that the gradient velocity vector field of ν_1 is $\boldsymbol{v} = \nabla_x x = \frac{d}{dx} x \equiv 1$, i.e. the Wasserstein gradient flow curve $P_{X,h}$ is the uniform location shift of the marginal distribution of X, and the scores of these counterfactual distributions are $\nabla_{\mathrm{K}} \nu_{1,P}(x) = -\frac{d}{dx} \varrho_X(x)/\varrho_X(x)$. The infinitesimal effects of the flow on an outcome functional $\psi(P_Y)$ follow from the von Mises formula (1.7) and integration by parts:

$$\begin{aligned}
\tfrac{d}{dh} \psi(P_h) &= \int \widetilde{\psi}_P(y) [-\tfrac{d}{dx} \varrho_X(x)/\varrho_X(x)] \, \mathrm{d}P(x, y) \\
&= \int_{\mathrm{supp(X)}} -\tfrac{d}{dx} \varrho_X(x) \, \mathrm{E}\Big[\widetilde{\psi}_P(Y)|X = x\Big] \, \mathrm{d}x \\
&= \int_{\mathrm{supp(X)}} \tfrac{d}{dx} \mathrm{E}\Big[\widetilde{\psi}_P(Y)|X = x\Big] \, \varrho_X(x) \mathrm{d}x.
\end{aligned}$$

Example 2.3 (Variance)**.** Continuing with the OB-setting of the previous example, we can obtain comparable counterfactual transport effects by changing any sufficiently regular functional of P_X. The variance functional $\nu_2(P_X) = \int (x - \nu_1(P))^2 dP_X(x)$ has influence function $\widetilde{\nu}_2(x) = (x - \nu_1(P))^2 - \nu_2(P)$ and gradient velocity vector field $\boldsymbol{v}_P = \nabla_x \widetilde{\nu}_{2,P}(x) = 2(x - \nu_1(P))$, which sends every agent away from the mean at the speed proportional to her distance from the mean. The effect of this perturbation on the variance is $\int_{\mathbb{R}} 2(x - \nu_1) \cdot 2(x - \nu_1) dP_X = 4\nu_2(P)$,

and the OB-effect of changing the variance of a covariate X along the Wasserstein gradient flow path $P_{X,h}$ on an outcome functional $\psi(P_Y)$ is

$$\tfrac{d}{dh}\psi(P_h) = \tfrac{1}{4\nu_2(P)} \int 2(x - \nu_1(P)) \; \tfrac{d}{dx}\mathrm{E}[\widetilde{\psi}_P(Y)|X = x] \; \mathrm{d}P_X.$$

Example 2.4 (Gaussian status quo)**.** Let $\varrho(x; \nu_1, \nu_2) = (2\pi\nu_2)^{-1/2} e^{-(x-\nu_1)^2/2\nu_2}$ denote the density of the Gaussian distribution with mean ν_1 and variance ν_2. It can be verified that the curve $\nu_1 \mapsto \varrho(\nu_1, \nu_2)$ has score $v_\varrho(x) = [x - \nu_1(\varrho)]/\nu_2(\varrho)$ that is equal to the nonparametric information gradient $\nabla^1_{\mathrm{F}}\nu_{1,\varrho}$ and to the Wasserstein gradient $\nabla^1_{\mathrm{K}}\nu_{1,\varrho}$ of the mean functional ν_1 at ϱ, and is therefore the gradient flow curve of ν_1 for any Gaussian status quo distribution in both metrics. Moreover, the curve $\nu_2 \mapsto \varrho(\nu_1, \nu_2)$ has score $u_\varrho(x) = [(x - \nu_1(\varrho))^2 - \nu_2(\varrho)]/2\nu_2^2$ that is equal to the information gradient $\nabla^1_{\mathrm{F}}\nu_{2,\varrho}$ and the Wasserstein gradient $\nabla^1_{\mathrm{K}}\nu_{2,\varrho}$ of the variance functional ν_2 at ϱ, and is therefore the gradient flow curve of ν_2 for any status quo Gaussian distribution in both metrics. Furthermore, these curves are orthogonal in both metrics.

Example 2.5 (Quantile)**.** Wasserstein gradient velocity vector field of the quantile functional

$$\nu_\tau(P_X) = \inf\{x \in \mathbb{R} \; ; \; F_X(x) \geq \tau\}$$

elucidates the reason for the density term $\varrho_X(\nu_\tau(P))$ in its asymptotic efficiency bound and influence function $\widetilde{\nu}_{\tau,P}(x) = (\tau - \mathbb{1}\{x \leq \nu_\tau(P)\})/\varrho(\nu_\tau(P))$. Because $\widetilde{\nu}_{\tau,P}$ is the Heaviside function, its derivative is the Dirac delta (generalized) function: The gradient velocity vector field of the quantile $\boldsymbol{v} = \nabla_x\widetilde{\nu}_\tau(x) = \delta_{\{\nu_\tau(P)\}}(x)/\varrho(\nu_\tau(P))$ transports *only* the agents at the quantile. The transportation cost of this velocity field, measured by the Wasserstein metric, is directly proportional to the mass or density at the quantile, therefore the derivative of the quantile functional is inversely proportional to the mass or density at the quantile. In the absolutely continuous case, gradient transport does not actually change ν_τ in a meaningful way, but the reweighting along the influence function (information gradient) as in eq. (2.12) is well-defined and its effect on ν_τ is inversely related to the density $\varrho(\nu_\tau)$.

The following companion result to lemma 7 shows the effect of reweighting the Wasserstein metric by a "cost distribution" term $\frac{dQ}{dP}$ on the direction of the gradient score.

Lemma 10 (Weighted Wasserstein gradient scores)**.** Suppose $\nu : \mathcal{P} \to \mathbb{R}$ is a sufficiently smooth functional, that P is a sufficiently regular probability measure, and the cost distributions satisfy $Q_P \ll P$ for all P with uniformly bounded derivatives $0 < m \leq \frac{dQ_P}{dP}(x) \leq M < \infty$ on $\mathcal{X}$. Then

$$\nabla_{\mathfrak{g}}\nu_P = -\mathrm{div}(\varrho\tfrac{dP}{dQ} \; \nabla_x\widetilde{\nu}_P)/\varrho \tag{2.23}$$

are the gradient scores for the weighted Wasserstein metric $\mathfrak{g}_{\mathrm{K}_Q,P}(\boldsymbol{v}, \boldsymbol{w}) = \int \langle \boldsymbol{v}, \boldsymbol{w} \rangle_{\mathbb{R}^d} \mathrm{d}Q_P$.

The interpretation of eq. (2.23) is that the Wasserstein metric charges a uniform price for moving an agent (unit of density or mass) x throughout the sample space $\mathcal{X}$. The effect of introducing the reweighting term $\frac{dQ_P}{dP}$ in the Wasserstein metric on the gradient score is to scale the velocity $\nabla_x\widetilde{\nu}_P(x)$ of transport of agents at location x by $\frac{dQ_P}{dP}^{-1}(x)$ to reflect the nonuniform cost of transport throughout the sample space.

3. Distributional counterfactuals

In Section 2 we have discussed how to specify a collection of scores $V = \{v_p\}_{p \in \mathcal{P}}$ on the tangent spaces $V(p) \in \mathcal{T}_p$ of the set of counterfactual distributions $\mathcal{P}$ with local optimality

conditions for policy changes in a given scalar functional $\nu : \mathcal{P} \to \mathbb{R}$. In this section[12], we construct and study an ordinary differential equation on the model of probability distributions $\mathcal{P}$, whose points are probability densities p, from the score field V:

$$\frac{\dot{p}}{p} = V(p) \in \mathcal{T}_p, \quad p(0) = p_0. \tag{3.1}$$

Equation (3.1) describes the construction of the one-dimensional model $\{p_h\}_{h\in J}$ of counterfactual distributions, that contains the status quo distribution p_0 and continuously evolves the current state p_h according to the direction $V(p_h)$ of the score field at that state, so that $\frac{d}{dh}\log p_h = V(p_h)$ for all $h \in J$. When the scores V are the (scaled) gradients of a functional ν, the path p_h is called the *gradient flow curve* of the functional ν. In this case equation (3.1) is a very useful algorithm for increasing the value of the functional by continuously changing the state p in the direction that most rapidly increases the value of ν.

3.1 Exponential statistical manifold. A formal treatment of this problem requires a manifold structure to be introduced on the collection of counterfactual distributions $\mathcal{P}$. Manifold structure of parametric families of probability distributions has been known in the statistical literature since at least Hotelling (1930) [43]. In the present paper we use the nonparametric parametrization that was introduced by Pistone and Sempi (1995) [61], who extend the work on finite-dimensional exponential families by Efron (1975) [29].

At each point $p \in \mathcal{P}$, we consider a set of densities of the form

$$q_u(x) = e^{u(x)-K_p(u)} \cdot p(x), \qquad K_p(u) = \log \mathrm{E}_p\big[e^{u(X)}\big], \tag{3.2}$$

where u is a score function at p and belongs to the neighborhood

$$\mathcal{S}_p \coloneqq \big\{u \in L^2_0(p)\ ;\ \mathrm{E}_p[e^u] < \infty\big\} \tag{3.3}$$

of 0 in a suitable Banach space $(B_p, \|\cdot\|_p)$ (defined below). Here $K_p(u)$ is a normalization constant. The mapping

$$B_p \supset \mathcal{S}_p \ni u \overset{\varphi_p}{\longmapsto} q_u \in \mathcal{P} \tag{3.4}$$

from the small scores u at p to the densities q_u near p in $\mathcal{P}$ is a nonparametric parametrization of a neighborhood $\varphi_p(\mathcal{S}_p)$ of p in $\mathcal{P}$ by the open subset $\mathcal{S}_p$ of a normed linear space $(B_p, \|\cdot\|_p)$. This should be compared to parametrization of a finite-dimensional statistical family by an open subset of some Euclidean space $(\mathbb{R}^d, |\cdot|)$. Unlike the case of a parametric model where a single mapping describes the entire model, the entire nonparametric model $\mathcal{P}$ is parametrized by the collection of maps $\{\varphi_p\}_{p\in\mathcal{P}}$. This poses the challenge to be certain that the same calculations performed in two different charts $\varphi_p(\mathcal{S}_p) \bigcap \varphi_q(\mathcal{S}_q)$ are consistent and provide the same results. We refer to Lang (1999) [48] for the general theory of infinite-dimensional manifolds modelled on Banach spaces. We refer to Pistone (2013) [60] for the technical details of the exponential manifold structure $\{\varphi_p\}$ of the nonparametric statistical model parametrized by (3.2).

The parametrization (3.2) of the model $\mathcal{P}$ by open subsets $\mathcal{S}_p$ of Banach spaces B_p identifies the tangent spaces $\mathcal{T}_p$ of the model with the linear spaces B_p. Specifically, the derivative mapping $\mathrm{d}\varphi_p$ is a bijection between B_p and $\mathcal{T}_p$. Recall that tangent spaces $\mathcal{T}_p$ and $\mathcal{T}_q$ at different points $p, q \in \mathcal{P}$ are distinct. This is because $\mathcal{P}$ is not a linear space. By contrast, the tangent spaces to the linear space B_p at any two vectors $u, v \in B_p$ is the linear space B_p itself, so are identical. Consequently, the parametrization (3.2) of $\mathcal{P}$ determines the mapping

$$B_q \ni u \mapsto u - \mathrm{E}_p[u] \in B_p \tag{3.5}$$

[12] New namespace, notation follows [60, 1].

of the tangent space $\mathcal{T}_q \equiv B_q$ at density $q \in \mathcal{P}$ onto the tangent space $\mathcal{T}_p \equiv B_p$ at density $p \in \mathcal{P}$, for densities q that are sufficiently close to p. This mapping of the tangent spaces allows the differential equation (3.1) to be formally defined and solved on an open subset of the single Banach space B_{p_0}:

$$p_{u(t)} = \frac{e^{u(t)}}{\mathrm{E}_{p_{u(t)}}\left[e^{u(t)}\right]} p_0, \tag{3.6}$$

$$\dot{u}(t) = \left\{ V(p_{u(t)}) - \mathrm{E}_{p_0}\left[V(p_{u(t)})\right] \right\} \in \mathcal{S}_{p_0} \subset B_{p_0}, \quad u(0) = 0. \tag{3.7}$$

Parametrization (3.2) also determines sufficient conditions for the existence and regularity of a solution to equation (3.7). Furthermore, the proof of the existence of a solution $J \ni t \mapsto u(t) \in B_{p_0}$ in eq. (3.7) is constructive and suggests a strategy for its numerical approximation and *estimation* of the counterfactual distributions p_t and their scalar effects $\Delta_h \psi(p_0)$ defined in eq. (0.2). We refer to Amann (1990) [1] and Lang (1999) [48] for the theory of ordinary differential equations in Banach spaces.

3.1.1 **Lipschitz condition** The Banach space used for parameterization of $\mathcal{P}$ in [61] is the Orlizc space $B_p = L_0^\Phi(p)$ of the Young function

$$\Phi(u) = \cosh u - 1 \tag{3.8}$$

of exponentially integrable functions with zero p-mean. A score function $u_p \in L_0^2(p)$ belongs to $L_0^{\cosh - 1}(p)$ if and only if the random variable $u_p(X)$ has a finite moment generating function $\mathrm{E}_p[e^{tu(X)}]$ in a neighborhood $t \in (-\epsilon, \epsilon)$ of zero. We note that this is a strong integrability condition, and remark that, to our knowledge, no alternative manifold parametrization of the nonparametric model is available in the mathematical and statistical literature.

The main technical condition that governs the differential eq. (3.7) involves the norm of the exponential Orlicz space $L^\Phi(p)$ defined by

$$\|u\|_{\Phi,p} := \inf\left\{ t > 0 \; ; \; \mathrm{E}_p\left[\Phi(u(X)/t)\right] < 1 \right\}. \tag{3.9}$$

A score field V satisfies the *local Lipschitz condition* at a density p if there exists a constant $\lambda > 0$ such that

$$\left\| \{V(q_u) - \mathrm{E}_p[V(q_u)]\} - \{V(q_v) - \mathrm{E}_p[V(q_v)]\} \right\|_{\Phi,p} \leq \lambda \|u - v\|_{\Phi,p} \tag{3.10}$$

for all scores u, v in a neighborhood of $0 \in L_0^\Phi(p)$. The open neighborhood in this condition is a subset of the proper domain $\mathcal{S}_p \subset L_0^\Phi(p)$ of the moment generating functional K_p.

3.2 Continuous exponential tilting (CET). We now use the Euler polygon scheme to show that under the Lipschitz regularity condition infinitesimal perturbations (scores) determine non-infinitesimal counterfactual distributions. We follow Amann (1990) [1, Chapter 2] closely in the development of the ODE existence theory.

Lemma 11 (Approximate solution via Euler polygon scheme)**.** Let $p_0 \in \mathcal{P}$. Suppose score field expressed V in local coordinates φ_{p_0}:

$$V = \left\{ v_q - \mathrm{E}_{p_0}[v_q] \; ; \; q \in \varphi_{p_0}\left(\mathcal{S}_{p_0}\right) \right\} \tag{3.11}$$

is Lipschitz on a neighborhood $D \subset \mathcal{S}_{p_0}$ of p_0. Let $b > 0$ such that the closed ball $\bar{B}(0, b) \subset D$. Let $M = \max_{u \in \bar{B}(0,b)} \|V(q_u)\|_{\Phi,p_0}$ and $\alpha = b/M$. Then for every $\epsilon > 0$ there exists an ϵ-approximate solution $u_\epsilon : [-\alpha, \alpha] \to \bar{B}(0, b)$ of the differential equation (3.7):

(i) $u_\epsilon \in C([-\alpha, \alpha], \bar{B}(0, b))$, and $u\epsilon$ is piecewise continuously differentiable;

(ii) For every subinterval $I \subset [-\alpha, \alpha]$ such that u_ϵ is continuously differentiable on I, we have $\|\dot{u}_\epsilon(t) - V(u_\epsilon(t))\|_{\Phi,p_0} \leq \epsilon$ for every $t \in I$;

(iii) We have the bound $\|u_\epsilon(t) - u_\epsilon(s)\|_{\Phi,p_0} \leq M|t-s|$ for all $t, s \in [-\alpha, \alpha]$.

Proof. Let $\delta = \epsilon/\lambda$, then by Lipschitz regularity

$$\|V(u) - V(v)\|_{\Phi,p_0} \leq \epsilon \quad \text{if} \quad \|u - v\|_{\Phi,p_0} \leq \delta.$$

Partition the interval $[-\alpha, \alpha]$ into subintervals

$$-\alpha =: t_{-n} < t_{-n+1} < \ldots < t_{-1} < t_0 = 0 < t_1 < \ldots < t_n := \alpha$$

such that $t_i - t_{i-1} \leq \min\{\delta, \delta/M\}$ for $i = -n+1, \ldots, n$.

Define

$$\begin{aligned} u_\epsilon(t) := &\sum_{i=1}^{n} \mathbb{1}_{\{t_{i-1} \leq t \leq t_i\}} \Big[\sum_{j=1}^{i-1} (t_j - t_{j-1}) V(u(t_{j-1})) + (t - t_i) V(u(t_{i-1})) \Big] \\ &+ \sum_{i=-n+1}^{0} \mathbb{1}_{\{t_{i-1} \leq t < t_i\}} \Big[\sum_{j=1}^{i-1} (t_j - t_{j-1}) V(u(t_j)) + (t - t_i) V(u(t_i)) \Big]. \end{aligned}$$

by choice of $\alpha = b/M$, u_ϵ is well-defined on $[-\alpha, \alpha]$ and remains in $\bar{B}(0, b)$. By construction, u_ϵ is continuous and satisfies (iii). Moreover, $\dot{u}_\epsilon(t) = V(u_\epsilon(t_i))$ for all $t \in [t_i, t_{i+1}] \cap [0, \infty)$ and $t \in [t_{i-1}, t_i] \cap (-\infty, 0]$ so that $\|u_\epsilon(t) - u_\epsilon(t_i)\| \leq \delta$ on the same intervals by the choice of the partition. Consequently (ii) holds by the choice of δ and Lipschitz property. □

The Euler polygon scheme is useful as a numerical algorithm for approximating the solution to eq. (3.7) and for establishing the existence of the solution theoretically. The next result provides a bound used to show that a sequence of approximate solutions u_ϵ converges to the unique solution of the equation as $\epsilon \to 0$.

Lemma 12. Let $p_0 \in \mathcal{P}$. Suppose score field expressed V in local coordinates φ_{p_0} as in eq. (3.11) and satisfies the Lipschitz condition (3.10) on neighborhood $D \subset \mathcal{S}_{p_0}$ of 0. If $u : J_u \to D$ and $v : J_v \to D$ are ϵ_1- and e_2-approximate solutions of $\dot{u} = V(u)$, then for every $t_0 \in J_u \cap J_v$ we have:

$$\|u(t) - v(t)\|_{\Phi,p_0} \leq \{\|u(t_0) - v(t_0)\|_{\Phi,p_0} + (\epsilon_1 + \epsilon_2)|t - t_0|\} e^{\lambda|t-t_0|} \tag{3.12}$$

for all $t \in J_u \cap J_v$.

Proof. First note that properties (i)-(ii) of ϵ-approximate solution in Lemma 11, it follows that

$$\begin{aligned} \left\| u(t) - u(t_0) - \int_{t_0}^{t} V(u(s))\, ds \right\|_{\Phi,p_0} &\leq \epsilon|t - t_0|, \quad \text{for all } t \in J_u \\ \left\| v(t) - v(t_0) - \int_{t_0}^{t} V(v(s))\, ds \right\|_{\Phi,p_0} &\leq \epsilon|t - t_0|, \quad \text{for all } t \in J_v. \end{aligned} \tag{3.13}$$

This follows by the fundamental theorem of calculus in the Banach space $L_0^\Phi(P_0)$ applied on each subinterval of continuous differentiability of u and v.

Next, using above estimates and the identity

$$u(t) - v(t) = \left[u(t) - u(t_0) - \int_{t_0}^{t} V(u(s))\, ds\right] - \left[v(t) - v(t_0) - \int_{t_0}^{t} V(v(s))\, ds\right] + [u(t_0) - v(t_0)] + \int_{t_0}^{t} \big[V(u(s)) - V(v(s))\big]\, ds,$$

we obtain by triangle inequality the bound

$$\|u(t) - v(t)\|_{\Phi,p_0} \le (\epsilon_1 + \epsilon_2)|t - t_0| + \|u(t_0) - v(t_0)\|_{\Phi,p_0} + \lambda \left| \int_{t_0}^{t} \|u(s) - v(s)\, ds\|_{\Phi,p_0} \right|$$

for all $t \in J_u \cap J_v$. The result follows from Gronwall's lemma. □

Lemma 12 implies that any two exact solutions $u : J_u \to \mathcal{S}_{p_0}$ and $v : J_v \to \mathcal{S}_{p_0}$ to the initial value problem (3.7) must coincide on $J_u \cap J_v$.

Theorem 13 (Local existance and uniquencess of counterfactual distributions)**.** Let $p_0 \in \mathcal{P}$. Assume that the score field V is expressed in local coordinates φ_{p_0} (3.2):

$$V = \Big\{ v_q - \mathrm{E}_{p_0}[v_q]\ ;\ q \in \varphi_{p_0}\big(\mathcal{S}_{p_0}\big) \Big\}$$

and is Lipschitz with constant λ (3.10) on a neighborhood $D \subset \mathcal{S}_{p_0}$ of 0. Let $b > 0$ such that the closed ball $\bar{B}(0,b) \subset D$. Let $M = \max_{u \in \bar{B}(0,b)} \|V(q_u)\|_{\Phi,p_0}$ and $\alpha = b/M$. Then the initial value problem (IVP)

$$p_{u(t)} = \frac{e^{u(t)}}{\mathrm{E}_{p_{u(t)}}\big[e^{u(t)}\big]} p_0,$$
$$\dot{u}(t) = V(p_{u(t)}), \quad u(0) = 0.$$

has a unique solution u on $[-\alpha, \alpha]$.

Proof. Fix a sequence $\epsilon_n \to 0$. By Lemma 11 there exist approximate solutions

$$u_{\epsilon_n} : [-\alpha, \alpha] \to \bar{B}(0,b), \qquad \text{for all } n \in \mathbb{N}$$

of the IVP. By Lemma 12, we have the following estimate

$$\|u_{\epsilon_n}(t) - u_{\epsilon_m}(t)\|_{\Phi,p_0} \le (\epsilon_m + \epsilon_m)\alpha e^{\lambda\alpha}, \quad \text{for all } t \in [-\alpha, \alpha]$$

for all $m, n \in \mathbb{N}$. Therefore sequence (u_{ϵ_n}) is Cauchy in the Banach space $C\big([-\alpha,\alpha], L^{\Phi}(p_0)\big)$, and converges uniformly on $[-\alpha, \alpha]$ to the function u, which must satisfy

$$u(t) = 0 + \int_0^t V(u(s))\, ds$$

by taking the limit in eq. (3.13). By the fundamental theorem of calculus, u is a solution to the IVP. The solution is unique by Lemma 12. □

We now consider simple special cases of CET that admit an explicit solution.

Example 3.1 (Gaussian OLS)**.** Consider the linear regression model with an outcome Y and a scalar covariate X that are jointly Gaussian:

$$Y = \beta_0 + \beta X + \epsilon, \quad \epsilon \perp\!\!\!\perp X, \quad \text{where } f_{XY,0} = dN\left(\begin{bmatrix} \chi_0 \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right). \tag{3.14}$$

In Example 2.4 it was shown that the information and Wasserstein gradient scores of the mean

functional

$$\chi(F) = \mathrm{E}_F[X], \quad \nabla_{\mathrm{F}}\chi_F(x) = \nabla_{\mathrm{K}}\chi_F(x) = x - \chi(F)$$

coincide at (marginal covariate) distributions F that are Gaussian. Moreover, changes along the influence function of the mean do not effect the variance $\sigma_X^2(F) = \mathrm{Var}_f[X]$ functional. It turns out that, if the initial distribution of the outcome and covariate is jointly Gaussian, then the solution to the ordinary differential equation (3.1) with the score field

$$V_F(x, y) = x - \chi(F)$$

is also given by a translation of the joint density that does not effect the covariance matrix:

$$f_{XY,t} = dN\left(\begin{bmatrix}\chi_0 + \sigma_X^2 t \\ \mu_0 + \rho\sigma_X\sigma_Y t\end{bmatrix}, \begin{bmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{bmatrix}\right).$$

In particular, the scalar counterfactuals $\Delta_t\mu = \mathrm{Cov}[X, Y]t$ and $\Delta_t\chi = \mathrm{Var}[X]t$ are linear in the time parameter t. Consequently, the counterfactual density and mean outcome, indexed by the change h in χ,

$$f_{XY,h} = dN\left(\begin{bmatrix}\chi_0 + h \\ \mu_0 + \beta t\end{bmatrix}, \begin{bmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{bmatrix}\right), \quad \Delta_{\chi,h}\mu = \mathrm{Cov}[X, Y]/\mathrm{Var}[X]h = \beta h,$$

follow from the linear time change $h_t = t/\mathrm{Var}[X]$. Apparently, the regression coefficient β measures the change in μ in terms of the change in χ exactly when the status quo is a Gaussian law. Also, $f_{XY,h}$ solves the differential equation (3.1) with the right-hand side $V_F(x, y) = [x - \chi(F)]/\sigma_X^2(F)$ rescaled by the information norm of the score.

We note that, despite a somewhat obscure definition via a functional ordinary differential equation, the statistical functionals $\Delta_t\mu(F_0)$ and $\Delta_t\chi(F_0)$ are regular parameters of the status quo distribution F_0. The example shows that these parameters are estimable at the parametric root-n rate on the nonparametric model when the true distribution is Gaussian. We expect this property to hold more generally (under regularity conditions). Another observation we make is that the estimator of the path of scalar counterfactuals

$$h \mapsto \widehat{\Delta_{\chi,h}\mu}(F_0) := \left\{\widehat{\mathrm{Cov}}[X, Y]/\widehat{\mathrm{Var}}[X] \cdot h\right\}_{0 \le h \le \epsilon}$$

has a uniform asymptotic distribution.

It can also be verified explicitly in this example that the fundamental theorem of calculus formula in equation (1.13) holds:

$$\begin{aligned}\Delta_{\chi,h}\mu =& \beta h = \int_0^h \beta(f_s)\,\mathrm{d}s \\ =& \int_0^t \int_{\mathbb{R}} \Big[\mathrm{E}[Y|X](x) - \mu(f_s)\Big]\Big[x - \chi(f_s)\Big] f_s(x)\mathrm{d}x\,\mathrm{d}s.\end{aligned}$$

This follows by using the standard fact that $\mathrm{E}[Y|X](x) = \rho\sigma_Y/\sigma_X(x - \chi_0) + \mu_0$ for a Gaussian distribution, the above formulas for $\mu(f_s)$ and $\chi(f_s)$, and the functional form of the normal density f_s.

The simplification afforded in eq. (3.1) by the Gaussian initial condition can also be used to solve explicitly for the distributional counterfactuals of the workhorse econometric technique that addresses the problem of endogeneity in structural equations models (SEMs).

Example 3.2 (Gaussian IV)**.** Consider the linear IV model of an outcome Y with a scalar

endogenous covariate X and a scalar instrument Z:

$$Y = \beta_0 + \beta_{\text{IV}} X + \varepsilon_Y$$
$$X = \gamma_0 + \gamma_1 Z + \varepsilon_X,$$

where $(Z, \varepsilon_X, \varepsilon_Y)$ are jointly Gaussian and disturbances $\varepsilon_X, \varepsilon_Y$ are independent of Z. It follows that (X, Y, Z) is a Gaussian random vector. The idea of the IV regression is that the parameter β_{IV} (defined by the moment condition $\varepsilon_Y \perp Z$) captures the relationship between the outcome Y and the covariate X that is not confounded by endogeneity. Our distributional (and nonparametric) interpretation of the IV coefficient β_{IV} is that it measures the effect of an (exogenous) change in the marginal distribution of the instrument Z that does not effect the conditional law $X, Y|Z$. Specifically, we interpret β_{IV} as the derivative of functional $\mu(F) = \mathrm{E}_F[Y]$ along the influence function $v_F(x, y, z) = z - \zeta(F)$ of $\zeta(F) = \mathrm{E}_F[Z]$, measured relative to the change in $\chi(F) = \mathrm{E}_F[X]$. It can be verified that the solution of the differential equation (3.1), determined by the score field v_F, with a Gaussian initial condition $f_{XYZ,0} = dN\Big([\chi_0, \mu_0, \zeta_0], \Sigma\Big)$, is the path of joint Gaussian distributions with a constant covariance matrix:

$$f_{XYZ,t} = dN\Big(\big[\chi_0 + \mathrm{Cov}[X, Z] \cdot t,\ \mu_0 + \mathrm{Cov}[Y, Z] \cdot t,\ \zeta_0 + \mathrm{Var}[Z] \cdot t\big],\ \Sigma\Big).$$

In particular, the changes in the mean parameters χ, μ, ζ are linear (in the time parameter t). It follows that the effect on the mean outcome, indexed by the change h in χ, is

$$\Delta_{\chi,h}\mu(F_0) = \mathrm{Cov}[Y, Z]/\,\mathrm{Cov}[X, Z] \cdot h = \beta_{\text{IV}} \cdot h,$$

by the linear reparametrization $h_t = t/\,\mathrm{Cov}[X, Z]$.

Example 3.3 (Non-Gaussian OLS)**.** Linearity of the mean outcome effect $h \mapsto \Delta_{\chi,h}\mu$ in Examples 3.1 and 3.2 is a consequence of orthogonality of the mean and variance functionals at Gaussian distributions. In general, these orthogonality and linearity properties do not hold.

However, the score flow equation (3.1) with the right-hand side given by the unscaled influence function of the mean $V_F(x, y) = x - \chi(F)$ corresponds to the flow of the constant vector field $\dot{u}(x, y) \equiv x - \chi(F_0)$ in the local coordinates at F_0 by (3.7). Consequently, this equation has an explicit solution given by the exponential tilting of the initial density:

$$f_t(x, y) = \frac{\exp\{(x - \chi_{F_0})t\}}{\mathrm{E}_{f_0}\big[\exp\{(x - \chi_{F_0})t\}\big]} f_0(x, y). \tag{3.15}$$

In the case of a non-Gaussian initial distribution F_0, the mean counterfactual effects

$$\Delta_t\chi(F_0) = \mathrm{E}_{f_0}[Xe^{Xt}]/\mathrm{E}_{f_0}[e^{Xt}] - \mathrm{E}_{f_0}[X], \quad \Delta_t\mu(F_0) = \mathrm{E}_{f_0}[Ye^{Xt}]/\mathrm{E}_{f_0}[e^{Xt}] - \mathrm{E}_{f_0}[Y] \tag{3.16}$$

need not be linear in the time parameter t along the integral path f_t.

Because f_t is the gradient flow of the mean χ, the map $t \mapsto \Delta_t\chi(F_0)$ is strictly increasing and smooth. Denote the functional determined by the inverse (with respect to time parameter t) of this map by

$$t_h(F_0) \coloneqq t\{h\}(F_0) \coloneqq \inf\{t\ ;\ \Delta_t\chi(F_0) \geq h\}. \tag{3.17}$$

We use parameter h to denote the change in the mean χ in contrast to the parameter t of the integral path (3.15) of the unscaled equation.

The solution f_h of the score flow equation (3.1) with the right-hand side given by the scaled gradient of the mean $V_F(x) = [x - \chi(F)]/\,\mathrm{Var}_F[X]$, where the parameter is $h = \Delta_h\chi(F_0)$ can be expressed as

$$f_h(x, y) = e^{(x - \chi(F_0)) \cdot t_h(F_0)}/\mathrm{E}_{f_0}\Big[e^{(X - \chi(F_0)) \cdot t_h(F_0)}\Big] \times f_0(x, y). \tag{3.18}$$

The corresponding change in the mean of the outcome can be expressed as

$$\Delta_{\chi,h}\mu(F_0) = \mathrm{E}_{f_0}\Big[Y e^{Xt_h(F_0)}\Big]/\mathrm{E}_{f_0}\Big[e^{Xt_h(F_0)}\Big] - \mathrm{E}_{f_0}[Y]. \tag{3.19}$$

We use the expression of $\Delta_{\chi,h}\mu$ in equation (3.19) to find the influence function $\widetilde{\Delta_h\mu}_{F_0}$ of this parameter of the status quo distribution F_0. Let F_s be a differentiable in quadratic mean path of distributions with a score v at F_0. By formula (2.3), the following equation is solved uniquely in $L_0^2(F_0)$ by the influence function:

$$\begin{aligned}\tfrac{d}{ds}_{|s=0}\Delta\mu(F_s) &= \mathrm{E}_{f_0}\Big[\widetilde{\Delta\mu}_{F_0}(X,Y)\, v(X,Y)\Big] \\ &= \tfrac{d}{ds}_{|s=0}\Big\{\mathrm{E}_{f_s}\big[Y e^{Xt_h(F_s)}\big]/\mathrm{E}_{f_s}\big[e^{Xt_h(F_s)}\big]\Big\} - \mathrm{E}_{f_0}\big[(Y-\mu(F_0))\, v(X,Y)\big]\end{aligned}$$

where we have dropped the subscripts χ, h for notational convenience. Using standard rules of differential calculus, and the Riesz representation of the pathwise derivatives of linear functionals (expectations) and the nonlinear time-change parameter $t_h(F)$, we find

$$\begin{aligned}\widetilde{\Delta\mu}_F(x,y) = &\Big\{y e^{xt_h(F)} - \mathrm{E}[Y e^{Xt_h(F)}]\Big\}/\mathrm{E}_f\Big[e^{Xt_h(F)}\Big] \\ &- \Big\{e^{xt_h(F)} - \mathrm{E}[e^{Xt_h(F)}]\Big\}/\mathrm{E}_f\Big[e^{Xt_h(F)}\Big]^2 - \big[y-\mu(F)\big] \\ &+ \Big\{\mathrm{E}_f\big[XY e^{Xt_h(F)}\big]/\mathrm{E}_f\big[e^{Xt_h(F)}\big] - \mathrm{E}_f\big[X e^{Xt_h(F)}\big]/\mathrm{E}_f\big[e^{Xt_h(F)}\big]^2\Big\} \times \widetilde{t\{h\}}_F(x).\end{aligned}$$

The first three terms of the influence function account for local variability in the functional when the time-change parameter $t_h(F)$ is fixed. The last term measures the local variability of the functional due to the variability in the time-change parameter.

The time-change functional $t\{h\}(F)$ is obtained from the unscaled effect $\Delta_t\chi(F)$ via an implicit function definition by requiring the relation $\Delta_{t\{h\}(F)}\chi(F) \equiv h$ to hold identically in F. Therefore, parameters $\Delta_t\chi(F)$ and $t\{h\}(F)$ share the same smoothness properties in the distributional argument F. This is similar to the pathwise differentiability properties of the distribution function evaluated at a point and the corresponding quantile functional. Specifically, we have

$$\widetilde{t\{h\}}_F(x) = \Big[-\big\{\tfrac{d}{dt}\Delta_t\chi(F)\big\}^{-1}\widetilde{\Delta_t\chi}_F(x)\Big]_{t=t_h(F)}.$$

Finally, the velocity $\frac{d}{dt}\Delta_t\chi(F)$ and the influence function $\widetilde{\Delta_t\chi}_F(x)$ can be computed (under regularity conditions) from the explicit definition of the effect $\Delta_t\chi(F)$ in eq. (3.16):

$$\begin{aligned}\tfrac{d}{dt}\Delta_t\chi(F) &= \mathrm{E}_f\big[X^2 e^{Xt}\big]/\mathrm{E}_f\big[e^{Xt}\big] - \mathrm{E}_f\big[X e^{Xt}\big]/\mathrm{E}_f\big[e^{Xt}\big]^2 \\ \widetilde{\Delta_t\chi}_F(x) &= \Big\{x e^{xt} - \mathrm{E}[X e^{Xt}]\Big\}/\mathrm{E}_f[e^{Xt}] \\ &\quad - \Big\{e^{xt} - \mathrm{E}[e^{Xt}]\Big\} \times \mathrm{E}_f[X e^{Xt}]/\mathrm{E}_f[e^{Xt}]^2 - \big\{x - \chi(F)\big\}.\end{aligned}$$

We conclude that the mean outcome effect functional $\Delta_{\chi,h}\mu$ is pathwise differentiable under regularity conditions.

Remark The influence function of a scalar effect $\Delta_t\psi$ along a general score flow path can be obtained by differentiating through the equation (3.1). See [1] for differentiability with respect to the initial condition of solutions to ODEs in Banach spaces. This requires taking the derivative of the influence function (higher order influence function, see Robins et al (2008) [63]). In econometrics, inference for parameters defined via a differential equation in Euclidean space has been obtained in Hausman and Newey (1995) [41] and Vanhems (2006) [73].

4. Appendix

4.1 Conditioning and scores of marginal distributions. A common situation encountered with Oaxaca-Blinder decomposition analysis is the paths of counterfactual distributions F_{XY} that hold the conditional distributions $F_{X|Y}$ fixed.

Proposition 14. A path of joint counterfactual distributions $h \mapsto F_{XY,h}$ has the Oaxaca-Blinder property that the conditional distributions $F_{Y|X,h} \equiv F_{Y|X}$ are constant if and only if the scores of this path $v_{f_h}(x,y) = \frac{d}{dh} \log f_{XY,h}(x,y) \equiv v_{f_h}(x)$ are constant in the outcome sample space variable y.

Moreover, if the conditional distributions $F_{Y|X}$ are constant along the path of joint distributions $F_{XY,h}$, then the score of the path of marginal distribution $F_{Y,h}$ is

$$\tfrac{d}{dh} f_{Y,h}(y) = \mathrm{E}_F\big[v_F(X)|Y=y\big]$$

is given by the conditional expectation of the joint score (which depends only on x).

Proof. For simplicity we assume that differentiation with respect to the parameter h can be done point-wise over the sample space $x, y \in \mathcal{X} \times \mathcal{Y}$. Suppose $f_{Y|X,h}$ does not depend on h. Using the definition of the conditional density function, the score of the joint path

$$\begin{aligned} v_f(x,y) &= \tfrac{d}{dh} \log f_{Y|X,h}(x,y) f_{X,h}(x) \\ &= \tfrac{d}{dh} \log f_{Y|X}(x,y) + \tfrac{d}{dh} \log f_{X,h}(x) \\ &= 0 + \tfrac{d}{dh} \log f_{X,h}(x) \end{aligned}$$

depends only on the sample space variable $x \in \mathcal{X}$.

Suppose the scores of the path $f_{XY,h}$ are $v_h(x,y) \equiv v_h(x)$ for all h. Then $\frac{d}{dh} f_{Y|X,h}(x,y)$ does not depend on h, and we can write the conditional probability density functions as

$$\log f_{Y|X,h} = A(x,h) + B(x,y) \quad \text{with} \quad e^{A(x,0)} \equiv 1.$$

For every t, we must have

$$1 \equiv \int e^{B(x,y)} e^{A(x,t)} \,\mathrm{d}y = e^{A(x,t)} \int e^{B(x,y)} \,\mathrm{d}y.$$

From the normalization at $h = 0$, it follows that $e^{B(x,y)} = f_{Y|X,0}(x,y)$. This implies that $e^{A(x,h)} = 1$ for all h and the conditional density does not depend on h.

Suppose $f_{XY,h}(x,y) = f_{Y|X}(y|x) f_{X,h}(x)$. We want to find the score of the path of marginal distributions $F_{Y,h}$. Assuming that we can exchange the order of integration and differentiation,

$$\begin{aligned} \tfrac{d}{dh} f_{Y,h}(y) &= \tfrac{d}{dh} \log \left\{ \int_{\mathcal{X}} f_{Y|X}(y|x) f_{X,h}(x) \,\mathrm{d}x \right\} \\ &= \frac{1}{f_{Y_h}(y)} \int_{\mathcal{X}} f_{Y|X}(y|x) \tfrac{d}{dh} f_{X,h}(h) \,\mathrm{d}x \\ &= \frac{1}{f_{Y_h}(y)} \int_{\mathcal{X}} \tfrac{d}{dh} \log f_{X,h}(x)\, f_{Y|X}(y|x) f_{X,h}(x) \,\mathrm{d}x \\ &= \frac{1}{f_{Y_h}(y)} \int_{\mathcal{X}} \tfrac{d}{dh} \log f_{X,h}(x)\, f_{XY}(x,y) \,\mathrm{d}x \\ &= \int_{\mathcal{X}} \tfrac{d}{dh} \log f_{X,h}(x)\, f_{X|Y}(x|y) \,\mathrm{d}x \\ &= \mathrm{E}_F\big[v_F(X)|Y=y\big] \end{aligned}$$

where $v_F(x) = \frac{d}{dh} \log f_{X,h}(x)$ is the score of the path of marginal distributions of X.

□

4.2 Influence function of overidentified GMM functional. We compute the influence function of the GMM parameter defined in eq. (1.23) on a fully nonparametric model, and record the required regularity conditions for pathwise differentiability of the functional.

Conditions G:

1. The functional $\theta_W(P)$ is well-defined and satisfies the first order condition of the minimization problem eq. (1.23) for each $P \in \mathcal{P}$;
2. Moment functions $g : \Theta \times \mathcal{X} \to \mathbb{R}^r$ are smooth in θ and integrable with respect to $\mathcal{P}$, the order of differentiation with respect to θ and integration over $\mathcal{X}$ with respect to $P \in \mathcal{P}$ can be interchanged;
3. The derivative $\partial_\theta g(\theta, x)$ smooth in θ and integrable in x with respect to $\mathcal{P}$, the order of differentiation in θ and integration with respect to $P \in \mathcal{P}$ can be interchanged;
4. The second derivative $\partial_\theta^2 g(\theta, x)$ is integrable;
5. Weighting matrix-valued functional is positive-definite and pathwise differentiable on $\mathcal{P}$ with influence function $\widetilde{W}_P$. This means that we have the Riesz representation of the derivative $dW_P(v) = \int_{\mathcal{X}} \widetilde{W}_P v \, \mathrm{d}P$ for all tangent vectors $v \in L_0^2(P)$.
6. The $p \times p$ matrix
$$\Big(\big(P[g(\theta_{W,P})]^T \, W_P \otimes I_p\big) \, P\Big[\partial_\theta \operatorname{vec}\big([\partial_\theta g(\theta_{W,P})]^T\big)\Big] + P\big[\partial_\theta g(\theta_{W,P})\big]^T W_P P\big[\partial_\theta g(\theta_{W,P})\big] \Big)$$
is nonsingular for every $P \in \mathcal{P}$.

Theorem 15 (Influence function in Proposition 3)**.** If conditions **G** hold, then the GMM functional $\theta_W(P)$ defined in eq. (1.23) is pathwise differentiable with the influence function $\widetilde{\theta}_{W,P}$ given in eq. (1.26).

Proof. Fix a path P_t in $\mathcal{P}$ that is differentiable in quadratic mean at P_0 with score function $\xi \in L_0^2(P_0)$. Let $\theta_t = \theta_W(P_t)$ denote the value of the functional along this path. Under the assumption that $\theta_W(P)$ is pathwise differentiable, the derivative along the path $\frac{d}{dt}_{t=0}\theta_t$ exists and is given by the action of the derivative mapping on the score function of the path $\frac{d}{dt}_{t=0}\theta_t = d\theta_{W,P_0}(\xi)$. All assumptions that must be made in order to compute $\frac{d}{dt}\theta_t$ are necessary conditions for pathwise differentiability.

The idea of the calculation of $\widetilde{\theta}_{W,P}$ is to find the Riesz representation of the pathwise derivative mapping $d\theta_{W,P}$ and recognize the influence function as the element of $L_0^2(P_0)$ that determines this representation. By Hilbert space theory, $\widetilde{\theta}_{W,P}$ is characterized uniquely by its inner products with all other elements of the tangent space $L_0^2(P_0)$. Since we allow the path P_t and its score $\xi \in L_0^2(P_0)$ to be arbitrary, the calculation fully characterizes the influence function $\widetilde{\theta}_{W,P_0}$. The Riesz representation theorem then implies pathwise differentiability under the conditions that are sufficient for the inner product representation of the derivative along an arbitrary path.

The GMM functional is characterized locally in a neighborhood of P_0 by the first order condition of the minimization problem (1.23) that defines it. Under the condition G1 this first order condition can be written as:
$$0 = P_t[\partial_\theta g(\theta_t)]^T \, W_t \, P_t[g(\theta_t)]. \tag{4.1}$$

We now apply the standard rules of classical calculus to differentiate (4.1) with respect to the parameter t of the path P_t:
$$0 = \underbrace{\tfrac{d}{dt} P_t[\partial_\theta g(\theta_t)]^T \, W_t \, P_t[g(\theta_t)]}_{=:T_G} + \underbrace{[\partial_\theta g(\theta_t)]^T \, \tfrac{d}{dt} P_t W_t \, P_t[g(\theta_t)]}_{=:T_W} + \underbrace{[\partial_\theta g(\theta_t)]^T \, W_t \, \tfrac{d}{dt} P_t P_t[g(\theta_t)]}_{=:T_g}$$

The three terms T_G, T_W, T_g defined above each yield a contribution to the influence function of the GMM functional. On the restricted model by the requirement that moment conditions (1.24) hold, the first two terms T_G, T_W that account for local changes in the second derivative of the moment function g and the weighting matrix W vanish.

We now compute each of the three terms.

Term T_G. We use denominator layout for derivatives of vectors (so that $\partial_\theta g(\theta)$ is an array of dimension $r \times p$. We refer to [24] for the details of matrix calculus. We use the vectorization identity

$$\operatorname{vec}(\mathbf{AXB}) = (\mathbf{B}^T \otimes \mathbf{A}) \operatorname{vec} \mathbf{X},$$

with $\mathbf{A} = I_p$, $\mathbf{X} = P_t[\partial_\theta g(\theta_t)]$, and $\mathbf{B} = P_0[g(\theta)]^T\, W_0$ in order to keep the equation expressed as a two-dimensional array after we apply chain rule and differentiate $\partial_\theta g(\theta_t)$ with respect to the vector θ:

$$\begin{aligned}\underset{p\times 1}{T_G} &= \tfrac{d}{dt}_{|t=0} P_t\big[\underbrace{\partial_\theta g(\theta_t)}_{r\times p}\big]^T\, W_0\; P_0[g(\theta_0)] \\ &= \underbrace{\Big(P_0[g(\theta_0)]^T\, W_0 \otimes I_p\Big)}_{:=T_{G,I},\quad p\times pr}\; \underbrace{\tfrac{d}{dt}_{|t=0} \operatorname{vec} P_t[\partial_\theta g(\theta_t)]^T}_{:=T_{G,II},\quad pr\times 1}\end{aligned}$$

The term $T_{G,I}$ is zero when the moment conditions (1.24) hold. The term $T_{G,II}$ is an array of dimension $pr \times 1$ and has the format

$$T_{G,II} = \begin{bmatrix} (P_t[\partial_{\theta_1} g_1(\theta_t)], \dots, P_t[\partial_{\theta_p} g_1(\theta_t)])^T \\ \vdots \\ (P_t[\partial_{\theta_1} g_r(\theta_t)], \dots, P_t[\partial_{\theta_p} g_r(\theta_t)])^T \end{bmatrix}$$

Consider the derivative with respect to t of the first element of the term $T_{G,II}$, applying chain rule and using the regularity condition G3, we have

$$\begin{aligned}(T_{G,II})_1 &= \tfrac{d}{dt} P_t[\partial_{\theta_1} g_1(\theta_t)] \\ &= P_0\Big[\tfrac{d}{dt}_{|t=0} \partial_{\theta_1} g_1(\theta_t)\Big] + \tfrac{d}{dt}_{|t=0} P_t[\partial_{\theta_1} g_1(\theta_t)] \\ &= P_0\Big[\partial_\theta \partial_{\theta_1} g_1(\theta_t) \cdot \tfrac{d}{dt}_{|t=0} \theta_t\Big] + \tfrac{d}{dt}_{|t=0} P_t[\partial_{\theta_1} g_1(\theta_0)] \end{aligned} \tag{4.2}$$

Under the hypothesis that $\theta_W(P)$ is pathwise differentiable, the derivative $\frac{d}{dt}_{|t=0}\theta_t$ has the Riesz representation:

$$\tfrac{d}{dt}_{|t=0}\theta_t = \tfrac{d}{dt}_{|t=0}\theta(P_t) = d\theta_{W,P_0}(\xi) = \int_{\mathcal{X}} \widetilde{\theta}_{W,P_0}(x)\, \xi(x)\, \mathrm{d}P_0 = P_0\Big[\widetilde{\theta}_{W,P_0}\xi\Big] \tag{4.3}$$

The strategy of our calculation is to differentiate $\frac{d}{dt}$ in eq. (4.1), which results in an expression that relates the influence function $\widetilde{\theta}_{W,P}$ and other terms that involve the primitives of the problem (moments g, their derivatives, weighting matrix W, etc). The point is to then solve for the influence function $\widetilde{\theta}_{W,P}$ in terms of the primitives of the problem.

The last term in eq. (4.2) is the derivative of the mean functional of the moment function given by $\partial_{th_1} g_1(\theta_0)$, it has the following Riesz representation of the derivative:

$$\begin{aligned}\tfrac{d}{dt}_{|t=0} P_t[\partial_{\theta_1} g_1(\theta_0)] &= \tfrac{d}{dt}_{|t=0} \int_{\mathcal{X}} \partial_{\theta_1} g_1(\theta_0)\, \mathrm{d}P_t \\ &= \int_{\mathcal{X}} \Big(\partial_{\theta_1} g_1(\theta_0)(x) - P_0[\partial_{\theta_1} g_1(\theta_0)]\Big)\, \xi(x)\, \mathrm{d}P_0 \\ &= P_0\Big[\big(\partial_{\theta_1} g_1(\theta_0) - P_0[\partial_{\theta_1} g_1(\theta_0)]\big)\, \xi\Big] \end{aligned} \tag{4.4}$$

We thus have the following expression for the term $T_{G,II}$:

$$T_{G,II} = \underbrace{P_0\Big[\partial_\theta \operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big]}_{pr\times p}\underbrace{P_0\Big[\widetilde{\theta}_{W,P_0}\cdot\xi\Big]}_{p\times 1} + \underbrace{P_0\Big[\Big(\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big) - P_0\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big)\cdot\xi\Big]}_{pr\times 1}.$$

We have the following expression for the term T_G:

$$\begin{aligned} T_G = &\underbrace{T_{G,I}P_0\Big[\partial_\theta \operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big]}_{p\times p} P_0\big[\widetilde{\theta}_{W,P_0}\cdot\xi\big] \\ &+ T_{G,I}P_0\Big[\Big(\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big) - P_0\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big)\cdot\xi\Big]. \end{aligned}$$

Term T_W. Calculation of this term is straightforward, and follows by condition G2 and linearity of expectation:

$$\begin{aligned} T_W &= P_0[\partial_\theta g(\theta_0)]^T\ \tfrac{d}{dt}W_t\ P[g(\theta_0)] \\ &= P_0[\partial_\theta g(\theta_0)]^T\ P_0[\widetilde{W}_{P_0}\cdot\xi]\ P_0[g(\theta_0)] \\ &= P_0\Big[P_0[\partial_\theta g(\theta_0)]^T\ \widetilde{W}_{P_0}\ P_0[g(\theta_0)]\cdot\xi\Big] \end{aligned}$$

Term T_g.

$$T_g = \underbrace{P_0[\partial_\theta g(\theta_0)]^T\ W_0}_{:=T_{g,I},\quad p\times r}\ \underbrace{\tfrac{d}{dt}_{|t=0}P_t[g(\theta_t)]}_{:=T_{g,II},\quad r\times 1}$$

To compute the term $T_{g,II}$ we apply chain rule, condition G1 and Riesz representation of the mean functional with moment function $g(\theta_0)$:

$$\begin{aligned} T_{g,II} &= \tfrac{d}{dt}_{|t=0}P_t\big[g(\theta_0)\big] + P_0\big[\tfrac{d}{dt}_{|t=0}g(\theta_t)\big] \\ &= P_0\Big[\big(g(\theta_0) - P_0[g(\theta_0)]\big)\cdot\xi\Big] + P_0\Big[\partial_\theta g(\theta_0)\Big]\ \tfrac{d}{dt}_{|t=0}\theta_W(P_t) \\ &= P_0\Big[\big(g(\theta_0) - P_0[g(\theta_0)]\big)\cdot\xi\Big] + P_0\Big[\partial_\theta g(\theta_0)\Big]\ P_0[\widetilde{\theta}_W\cdot\xi] \end{aligned}$$

We have the following expression for the term T_g:

$$T_g = T_{g,I}P_0\Big[\big(g(\theta_0) - P_0[g(\theta_0)]\big)\cdot\xi\Big] + \underbrace{T_{g,I}P_0\Big[\partial_\theta g(\theta_0)\Big]}_{p\times p}\ P_0[\widetilde{\theta}_{W,P_0}\cdot\xi]$$

Collection all three calculations together we obtain:

$$\begin{aligned} &-\underbrace{\Big(T_{G,I}P_0\Big[\partial_\theta \operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big] + T_{g,I}P_0\Big[\partial_\theta g(\theta_0)\Big]\ \Big)}_{p\times p}P_0\Big[\widetilde{\theta}_{W,P_0}\cdot\xi\Big] = \\ &\quad P_0\Bigg[\Big\{T_{G,I}\Big(\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big) - P_0\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big) + \\ &\qquad + P_0[\partial_\theta g(\theta_0)]^T\ \widetilde{W}_{P_0}\ P_0[g(\theta_0)] + T_{g,I}\big(g(\theta_0) - P_0[g(\theta_0)]\big)\Big\}\cdot\xi\Bigg] \end{aligned}$$

Using condition G6, we solve for the Riesz representation of $d\theta_{W,P_0}$:

$$P_0\Big[\widetilde{\theta}_{W,P_0}\cdot\xi\Big] = -\Big(T_{G,I}P_0\Big[\partial_\theta \operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big] + T_{g,I}P_0\Big[\partial_\theta g(\theta_0)\Big]\Big)^{-1}\times$$
$$\times P_0\Bigg[\Big\{T_{G,I}\Big(\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big) - P_0\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big)+$$
$$+ P_0[\partial_\theta g(\theta_0)]^T\,\widetilde{W}_{P_0}\,P_0[g(\theta_0)] + T_{g,I}\big(g(\theta_0) - P_0[g(\theta_0)]\big)\Big\}\cdot\xi\Bigg]$$

Since the path P_t and its score $\xi \in L_0^2(P_0)$ are arbitrary, we conclude that

$$\widetilde{\theta}_{W,P_0} = -\Big[T_{G,I}P_0\Big[\partial_\theta \operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big] + T_{g,I}P_0\Big[\partial_\theta g(\theta_0)\Big]\Big]^{-1}$$
$$\times\Big\{T_{G,I}\Big(\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big) - P_0\operatorname{vec}\big([\partial_\theta g(\theta_0)]^T\big)\Big)+$$
$$+ P_0[\partial_\theta g(\theta_0)]^T\,\widetilde{W}_{P_0}\,P_0[g(\theta_0)] + T_{g,I}\big(g(\theta_0) - P_0[g(\theta_0)]\big)\Big\}$$

Substituting the terms $T_{G,I}$ and $T_{g,I}$, we obtain the expression for the nonparametric influence function of the GMM functional:

$$\widetilde{\theta}_{W,P} = -\Big[\big(P[g(\theta_P)]^T W_P\otimes I_p\big)\,P\big[\partial_\theta \operatorname{vec}\big([\partial_\theta g(\theta_P)]^T\big)\big] + P[\partial_\theta g(\theta_P)]^T W_P P\big[\partial_\theta g(\theta_P)\big]\Big]^{-1}$$
$$\times\Big\{\Big(P[g(\theta_P)]^T W_P\otimes I_p\Big)\Big[\operatorname{vec}\big([\partial_\theta g(\theta_P)]^T\big) - P\operatorname{vec}\big([\partial_\theta g(\theta_P)]^T\big)\Big]+$$
$$+ P\big[\partial_\theta g(\theta(P))\big]^T\,\widetilde{W}_P\,P\big[g(\theta_P)\big] + P\big[\partial_\theta g(\theta_P)\big]^T W_P\Big(g(\theta_P) - P\big[g(\theta_P)\big]\Big)\Big\}.$$

□

4.3 Gradient scores.

Proof of Proposition 4 . By the Riesz representation theorem for the Hilbert space $(\mathfrak{T}_P, \mathfrak{g}_P)$, applied to the pathwise derivative $\mathrm{d}\nu_P$ has the representation and by the Cauchy-Schwarz inequality:

$$\mathrm{d}\nu_P(v_P) = \mathfrak{g}_P(\nabla_{\mathfrak{g}}\nu_P, v_P) \le \|\nabla_{\mathfrak{g}}\nu_P\|_{\mathfrak{g},P}\|v_P\|_{\mathfrak{g},P}$$

with equality holding if and only if v_P is a scalar multiple of the gradient $\nabla_{\mathfrak{g}}\nu_P$. This implies a lower bound on the size of the score function v_P that leads to a unitary perturbation $\mathrm{d}\nu_P(v_P) = 1$ in the value of the functional:

$$1/\|\nabla_{\mathfrak{g}}\nu_P\|_{\mathfrak{g},P} \le \|v_P\|_{\mathfrak{g},P} \quad \text{for all} \quad v_P\in\mathfrak{T}_P \quad \text{s.t.} \quad \mathrm{d}\nu_P(v_P)\ge 1.$$

By the Cauchy-Schwarz inequality, the lower bound is achieved uniquely by the score function $\nabla^1_{\mathfrak{g}}\nu_P = \nabla_{\mathfrak{g}}\nu_P/\mathrm{Cov}\big[\widetilde{\nu}_P, \nabla_{\mathfrak{g}}\nu_P\big]$, which is a scalar multiple of the gradient and has the unitary local effect on the functional:

$$\mathrm{d}\nu_P(\nabla^1_{\mathfrak{g}}\nu_P) = \int_{\mathcal{X}}\widetilde{\nu}_P\,\nabla^1_{\mathfrak{g}}\nu_P\,\mathrm{d}P = \mathrm{Cov}_P\big(\widetilde{\nu}_P, \nabla^1_{\mathfrak{g}}\nu_P\big) = 1.$$

□

4.4 Influence function formula. The example below shows that the original von Mises (1947) calculation relies on paths that are not smooth in the sense required for pathwise differentiability.

Example 4.1 (Influence function for the von Mises calculation)**.** Let P, δ_x be a continuous distribution and a point mass on $(\mathcal{X}, \mathcal{A})$ and take $[0,1] \ni h \mapsto P_h = (1-h)P + h\delta_x$ to be a deviation as in the original von Mises calculation of influence function described in Section 1.1. We compute the tangent vector to the path P_h at $h = 1/3$ and $h = 0$. Take $\mu = P + \delta_x$ to be the dominating measure, then $dP_h = (1-h)\mathbb{1}_{\{\mathcal{X}\setminus x\}} + h\mathbb{1}_{\{x\}}$ is the density and $dP_h^{1/2} = \sqrt{1-h}\mathbb{1}_{\{\mathcal{X}\setminus x\}} + \sqrt{h}\mathbb{1}_{\{x\}}$ is the geometric representation of P_h under embedding in H_2. Note that the path is no longer linear. Also note that $d_{\mathrm{TV}}(P_{h+t}, P_h) = 2\sup_{A\in\mathcal{A}}|P_{h+t}(A) - P_h(A)| = t$, so the path is continuous in H_2. For $h = 1/3$, we can differentiate $\frac{d}{dt}|_{t=0} \log dP_{h+t}^{1/2}$ pointwise for each $z \in \mathcal{X}$ and verify that the score function $\frac{1}{2}v_{\frac{1}{3}}(z)dP_{\frac{1}{3}}^{1/2}(z) = -\frac{1}{2}(\frac{2}{3})^{-1/2}\mathbb{1}_{\{\mathcal{X}\setminus x\}}(z) + \frac{1}{2}(\frac{1}{3})^{-1/2}\mathbb{1}_{\{x\}}(z)$ satisfies the smoothness condition (2.1):

$$\begin{aligned}
&\left\| t^{-1}\left[dP_{\frac{1}{3}+t}^{-1/2} - dP_{\frac{1}{3}}^{-1/2}\right] - \tfrac{1}{2}v_{\frac{1}{3}}dP_{\frac{1}{3}}^{-1/2} \right\|_{H_2}^2 \\
&\quad = \left[t^{-1}[\sqrt{\tfrac{2}{3}-t} - \sqrt{\tfrac{2}{3}}] - (-1)\frac{1}{2}(\frac{2}{3})^{-1/2}\right]^2 P(\mathcal{X}\setminus x) + \left[t^{-1}[\sqrt{\tfrac{1}{3}+t} - \sqrt{\tfrac{1}{3}}] - \frac{1}{2}(\frac{1}{3})^{-1/2}\right]^2 \delta_{\{x\}}(x)
\end{aligned}$$

is $o(1)$ as $t \to 0$. For $h = 0$ the right derivative of $\sqrt{h}$ is infinite, so there is no score function v with finite values a.e.-μ that would satisfy condition (2.1). Consequently, the path P_h is not smooth in H_2 and *does not* have a tangent vector at $h = 0$, despite being a convex (linear) combination of probability measures.

Lemma 16 (Approximation to von Mises perturbation with a score)**.** Suppose K is a bounded probability density function on $\mathbb{R}^d$ with support in the unit ball $|x| \le 1$. Then

$$K_\delta(x) := \delta^{-d}K(\delta^{-1}x), \quad \delta > 0 \tag{4.5}$$

is an *approximation to the identity* in the sense of Stein and Shakarchi (2009) [67, p. 109], that is

(i) $\int_{\mathbb{R}^d} K_\delta(x)\,dx = 1$.

(ii) $|K_\delta(x)| \le A\delta^{-d}$ for all $\delta > 0$.

(iii) $|K_\delta(x)| \le A\delta/|x|^{d+1}$ for all $\delta > 0$ and $x \in \mathbb{R}^d$.

Here A is a constant independent of δ.

Suppose P_0 is a probability measure that is absolutely continuous with respect to the Lebesgue measure $\mathscr{L}^d$ with a continuous density function f_0. Let

$$K_{f_0,\delta,z}(x) := \left[\int_{\{f_0>\delta\}} K_\delta(x)\,dx\right]^{-1} \mathbb{1}_{\{f_0>\delta\}}(z-x)K_\delta(z-x), \tag{4.6}$$

then for $z \in \{f_0 > 0\}$ we have $K_{f_0,\delta,z} = K_\delta(z - .)$ for all sufficiently small $\delta > 0$ (which depend on z that is fixed throughout). Furthermore,

$$f_{t,\delta,z}(x) := (1-t)f_0(x) + tK_{f_0,\delta,z}(x) \tag{4.7}$$

is a curve of probability densities for t in an interval around 0, that is differentiable in quadratic mean of eq. (2.1) with the score function

$$v_{\delta,z}(x) := \tfrac{d}{dt}|_{t=0} \log f_{t,\delta,z}(x) = \frac{K_{f_0,\delta,z}(x)}{f_0(x)} - 1. \tag{4.8}$$

Proof. The three properties of an approximation to the identity follow respectively from dilation invariance of Lebesgue integral, boundedness and compact support of the kernel K.

Fix a $z \in \{f_0 > 0\}$. By continuity of f_0 there is a neighborhood $\mathcal{N}$ of z such that f_0 is bounded away from zero on $\mathcal{N}$. For all $\delta > 0$ small enough, $K_\delta(z - .)$ is supported in $\mathcal{N}$ by

bounded support and dilation construction, so that $K_\delta(z-.) \equiv K_{f_0,\delta,z}$. Therefore for t negative and close enough to 0, function $f_{t,\delta,z}$ is a well-defined probability density and its score functions

$$v_{t,\delta,z}(x) \coloneqq \tfrac{d}{dt} \log f_{t,\delta,z}(x) = \frac{K_{f_0,\delta,z}(x) - f_0(x)}{f_{t,\delta,z}(x)} \tag{4.9}$$

are bounded in $x \in \mathcal{X}$.

To check the DQM property

$$\int_{\mathcal{X}} \left[\frac{\sqrt{f_{t,d,z}} - \sqrt{f_0}}{t} - \frac{1}{2} v_{0,\delta,z} \sqrt{f_0} \right]^2 dx \to 0 \quad \text{as } t \to 0, \tag{4.10}$$

note that the map $t \mapsto \sqrt{f_{t,\delta,z}}(x)$ is continuously differentiable for each x in a neighborhood $t \in (-\epsilon, \epsilon)$ of 0 with derivative $\frac{1}{2} v_{t,\delta,z}(x) \sqrt{f_{t,\delta,z}(x)}$, therefore the problem is to justify the change of order of the limit $t \to 0$ and the integral $\int_{\mathcal{X}} dx$ in eq. (4.10). By the fundamental theorem of calculus, we can write the difference quotient as

$$\sqrt{f_{0+ht}(x)} - \sqrt{f_0(x)} = \int_0^1 \frac{d}{dh} \sqrt{f_{0+ht}(x)}\, dh = \int_0^1 \frac{1}{2} v_{0+ht}(x) \sqrt{f_{0+ht}} \cdot t \, dh.$$

Therefore, by $(a-b)^2 \le 2a^2 + 2b^2$ and Cauchy-Schwarz inequality, we have the pointwise bound

$$\begin{aligned} \left[\frac{\sqrt{f_{t,d,z}(x)} - \sqrt{f_0}(x)}{t} - \frac{1}{2} v_{0,\delta,z}(x) \sqrt{f_0(x)} \right]^2 &\le 2 \left[\int_0^1 \frac{1}{2} v_{ht,\delta,z}(x) \sqrt{f_{ht,\delta,z}}\, dh \right]^2 + 2 \frac{1}{2} v_{\delta,z}(x)^2 f_0(x) \\ &\le \int_0^1 \frac{1}{2} v_{ht,\delta,z}(x)^2 f_{0+ht}\, dh \; + v_{\delta,z}(x)^2 f_0(x). \end{aligned}$$

By the generalized Lebesgue dominated convergence theorem [see 64, p89, Theorem 19], in order to conclude (4.10), it is sufficient to show that $\int_{\mathcal{X}} \int_0^1 \frac{1}{2} v_{ht,\delta,z}(x)^2 f_{0+ht}\, dh dx$ converges as $t \to 0$. By Fubini's theorem

$$\int_{\mathcal{X}} \int_0^1 \frac{1}{2} v_{ht,\delta,z}(x)^2 f_{0+ht}\, dh dx = \int_0^1 \int_{\mathcal{X}} \frac{1}{2} v_{ht,\delta,z}(x)^2 f_{0+ht}\, dx dh = \frac{1}{2} \int_0^1 I_{ht,\delta,z}\, dh.$$

Since the scores (4.9) are bounded, the information matrix $I_{t,\delta,z}$ is continuous in t at 0, and the above integral converges to $I_{0,\delta,z}$. □

Proof of Theorem 8. By pathwise differentiability of functional ψ, differentiability in quadratic mean of the path $t \to P_{t,\delta,z}$, and Riesz representation we have

$$\begin{aligned} \tfrac{d}{dt}_{|t=0} \psi(P_{t,\delta,z}) &= \mathrm{d}\psi_{P_0}(v_{\delta,z}) \\ &= \int \widetilde{\psi}_{P_0}(x)\, v_{\delta,z}(x)\, dP_0. \end{aligned}$$

Assume that $\widetilde{\psi}_{P_0}(x) = \psi_{P_0}(x) \mathbb{1}_{\{f_0 > 0\}}(x)$. Below P_0 is fixed and we drop the subscript P_0 for convenience. Using the score $v_{\delta,z}$ computed in Lemma 16 and the fact that $\widetilde{\psi}$ has zero mean, have the expression for pathwise derivative as the convolution of influence function with the approximation to identity kernels:

$$\begin{aligned} \tfrac{d}{dt}_{|t=0} \psi(P_{t,\delta,z}) &= \int \widetilde{\psi}_{P_0}(x) \left[K_\delta(z-x)/f_0(x) - 1 \right] dP_0 \\ &= \int \widetilde{\psi}_{P_0}(x)\, K_\delta(z-x)\, dx \\ &= (\widetilde{\psi}_{P_0} * K_\delta)(z). \end{aligned}$$

It suffices to show that for each $\alpha > 0$ the set

$$E_\alpha = \left\{ z \in \mathrm{supp} P_0 \; ; \; \limsup_{\delta \to 0} \left| (\widetilde{\psi}_{P_0} * K_\delta)(z) - \widetilde{\psi}_{P_0}(z) \right| > 2\alpha \right\}$$

has zero Lebesgue measure, because then $E = \bigcup_{j=1}^{\infty} E_{1/j}$ has zero measure by monotonicity, and the assertion eq. (2.17) of the Theorem holds at all points $z \in E^c$ of the complement.

Because K_δ is a bounded probability density function, with support in $|x| \leq \delta$ by the dilation construction (4.5), we can write

$$\begin{aligned}
\left| (\widetilde{\psi}_{P_0} * K_\delta)(z) - \widetilde{\psi}_{P_0}(z) \right| &= \left| \int_{\mathbb{R}^d} \left[\widetilde{\psi}_{P_0}(z-x) - \widetilde{\psi}_{P_0}(z) \right] K_\delta(x) \, dx \right| \\
&\leq \int_{\mathbb{R}^d} \left| \widetilde{\psi}_{P_0}(z-x) - \widetilde{\psi}_{P_0}(z) \right| K_\delta(x) \, dx \\
&\leq \frac{c}{\delta^d} \int_{|x| \leq \delta} \left| \widetilde{\psi}_{P_0}(z-x) - \widetilde{\psi}_{P_0}(z) \right| dx.
\end{aligned}$$

Fix $\alpha > 0$ and recall that continuous functions of compact support are dense in $L^1(\mathbb{R}^d)$ [see e.g. 67, p71], so that for each $\epsilon > 0$ we can choose a function g with $\|\widetilde{\psi}_{P_0} - g\|_{L^1(\mathbb{R}^d)} < \epsilon$. By triangle inequality we can upper bound the expression above with

$$\frac{c}{\delta^d} \int_{|x| \leq \delta} \left| \widetilde{\psi}_{P_0}(z-x) - g(z-x) \right| dx + \frac{c}{\delta^d} \int_{|x| \leq \delta} \left| g(z-x) - g(z) \right| dx + c' |g(z) - \widetilde{\psi}_{P_0}(z)|.$$

By the continuity of g it follows that

$$\lim_{\delta \to 0} \frac{c}{\delta^d} \int_{|x| \leq \delta} \left| g(z-x) - g(z) \right| dx = 0, \qquad \text{for all } z.$$

We find that

$$\limsup_{\delta \to 0} \left| (\widetilde{\psi}_{P_0} * K_\delta)(z) - \widetilde{\psi}_{P_0}(z) \right| \leq c' \left| \widetilde{\psi}_{P_0} - g \right|^*(z) + c' \left| g(z) - \widetilde{\psi}_{P_0}(z) \right|,$$

where the superscript $*$ indicates the Hardy-Littlewood maximal function:

$$f^*(x) \coloneqq \sup_{B \ni x} \frac{1}{\mathscr{L}^d(B)} \int_B |f(y)| \, dy, \qquad \text{for } f \in L^1(\mathbb{R}^d), \quad x \in \mathbb{R}^d. \tag{4.11}$$

If we

$$F_\alpha = \{ z \in \mathrm{supp} P_0 \; ; \; \left| \widetilde{\psi}_{P_0} - g \right|^*(z) > \alpha \} \quad \text{and} \quad G_\alpha = \{ z \in \mathrm{supp} P_0 \; ; \; \left| \widetilde{\psi}_{P_0}(z) - g(z) \right| > \alpha \}$$

then $E_\alpha \subset F_\alpha \cup G_\alpha$ by De Morgan's law since $E_\alpha^c \supset F_\alpha^c \cap G_\alpha^c$. Furthermore, by Chebyshev's inequality

$$\mathscr{L}^d(G_\alpha) \leq \frac{1}{\alpha} \|\widetilde{\psi}_{P_0} - g\|_{L^1(\mathbb{R}^d)},$$

and by the Hardy-Littlewood maximal inequality [see e.g. 67, p101]

$$\mathscr{L}^d(F_\alpha) \leq \frac{3^d}{\alpha} \|\widetilde{\psi}_{P_0} - g\|_{L^1(\mathbb{R}^d)}.$$

Recall that the function g was chosen such that $\|\widetilde{\psi}_{P_0} - g\|_{L^1(\mathbb{R}^d)} < \epsilon$, so that

$$\mathscr{L}^d(E_\alpha) \leq c' \frac{3^d}{\alpha} \epsilon + c' \frac{1}{\alpha} \epsilon.$$

Since $\epsilon > 0$ is arbitrary, we conclude that $\mathscr{L}^d E_\alpha = 0$ and consequently $P_0(\bigcup_{j=1}^{\infty} E_{1/j}) = 0$. $\square$

4.5 Information counterfactuals of regular estimators. Here we derive a formula for the information gradient sensitivity and effects when both the policy and outcome functionals are specified implicitly with estimators.

Let data $X_1, \ldots, X_n$ be sampled randomly from $P \in \mathcal{P}$. We have shown that counterfactual values and locally optimal policy scores can be obtained analytically for statistical functionals that are smooth. However, empirical research in economics is typically formulated in terms of estimators rather than the statistical functionals that are estimated. Suppose we have two scalar estimators $\widehat{\psi}, \widehat{\nu} = \widehat{\psi}_n, \widehat{\nu}_n(X_1, ..., X_n)$. We think of the functionals $\psi, \nu : \mathcal{P} \to \mathbb{R}$ as the sets of large sample limits of the estimators $\widehat{\psi}, \widehat{\nu}$, indexed by the counterfactual distributions $P \in \mathcal{P}$ of data. Parameters $\psi(P_0)$ and $\psi(P_h)$ are the initial and the counterfactual limits of the outcome estimator $\widehat{\psi}$, computed with data sampled before and after the policy that changes the population value of ν by h. The innovation of this paper is to express the limit of $\widehat{\psi}$ at a carefully chosen counterfactual distribution P_h as a function (1.9) of the initial distribution P_0. This means that both parameters $\psi(P_0)$ and $\psi(P_h)$ can be estimated simultaneously from the same data.

Estimators are taken in a wide sense of any sequence of measurable transformations of the data with arbitrary dependence on the data and the sample size n. By contrast, von Mises considered simple estimators constructed directly from the functional by evaluating it at the empirical distribution. Consequently, the problems of asymptotic inference and counterfactual analysis can be approached either from the perspective of the calculus of the functional following von Mises, or from the perspective of a given estimator, as is typically done in econometrics. In order to compute counterfactual effects in statistical parameters with von Mises calculus, regularity conditions must be satisfied by the path of counterfactual distributions P_h, the functionals ψ, ν and the estimators $\widehat{\psi}, \widehat{\nu}$. The high level condition is that the paths and the functionals must be smooth. More primitive regularity conditions on $P \in \mathcal{P}$ can be formulated either based on the functionals or the estimators of the functionals. We record sufficient regularity conditions in terms of estimators of the policy and outcome parameters.

Asymptotic linearity (2.8) on $\mathcal{P}$ is the basic property of an estimator that parallels the Taylor expansion of the von Mises functional. Condition (2.8) pins down the parametric $\sqrt{n}$-rate of convergence of the estimator sequence to the functional, the asymptotic Gaussian distributions and candidates $\widetilde{\psi}_P$ for information gradients $\nabla_{\text{F}}\psi_P$, suggesting that the estimated functional must be smooth. Indeed, Donoho and Liu (1991) [26, 27] showed that parameters that are not pathwise differentiable require nonparametric estimators with rates of convergence strictly slower than $O_P(n^{-1/2})$. However, asymptotic linearity is a property that is determined by the behavior of the estimator at each fixed distribution P in isolation, without taking into account the behavior of the estimator at any other distribution. Strictly speaking, (2.8) is too weak for analytic counterfactual analysis, allowing for pathological behavior on a small set of counterfactual distributions. Consider the canonical example:

Example 4.2 (Hodges' estimator, hard thresholding)**.** Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\psi(P) = \int_{\mathcal{X}} x\, dP$ be the sample and the population means. Functional ψ is pathwise differentiable with information gradient $\widetilde{\psi}_P(x) = x - \psi(P)$. Note that $\widetilde{\psi}_P$ is the influence function of the estimator $\overline{X}_n$. Fix $a \in \mathbb{R}$ and let $\widehat{\psi}_{\text{H}} = \overline{X}_n \mathbb{1}_{\{|\overline{X}_n| \geq n^{-1/4}\}} + a\overline{X}_n \mathbb{1}_{\{|\overline{X}_n| < n^{-1/4}\}}$, then $\widehat{\psi}_{\text{H}}$ is asymptotically linear with influence function $\widetilde{\psi}_{\text{H},P} = [1 + (a-1)\mathbb{1}_{\{P\,;\,\psi(P)=0\}}]\widetilde{\psi}_P$ that is different from the information gradient of the functional at every distribution P with zero mean.

Proof. Case 1: $\psi_P \neq 0$. Write

$$\sqrt{n}(\widehat{\psi} - \psi_P) = \sqrt{n}(\overline{X}_n - \psi_P) + \big[\sqrt{n}(a-1)(\overline{X}_n - \psi_P) + \sqrt{n}(a-1)\psi_P\big]\mathbb{1}_{\{|\overline{X}_n| \leq n^{-1/4}\}}$$

and note that $P\{|\overline{X}_n| \leq n^{-1/4}\} = P\{|\sqrt{n}(\overline{X}_n - \psi_P) + \sqrt{n}\psi_P| \leq n^{1/4}\} \to 0$.

Case 2: $\psi_P = 0$. Write

$$\sqrt{n}(\widehat{\psi} - \psi_P) = \sqrt{n}a(\overline{X}_n - \psi_P) + \left[\sqrt{n}(1-a)(\overline{X}_n - \psi_P)\right]\mathbb{1}_{\{|\overline{X}_n| > n^{-1/4}\}}$$

and note that $P\{|\overline{X}_n| > n^{-1/4}\} = P\{|\sqrt{n}(\overline{X}_n - \psi_P)| > n^{1/4}\} \to 0$. Conclude that $\widehat{\psi}_H$ is asymptotically equivalent to the sample mean $\overline{X}_n$ on $\{P\,;\ \psi_P \neq 0\}$ and to $a\overline{X}_n$ on $\{P\,;\ \psi_P = 0\}$ and is therefore asymptotically linear at any distribution with finite variance. □

Example 4.2 shows that asymptotic linearity at P is not enough to find the information gradient of the functional, and therefore does not alone provide information about the behavior of the estimator at counterfactual distributions. An estimator $\widehat{\psi}_n$ is *locally regular* at P_0, if for every smooth in the sense of differentiability in quadratic mean eq. (2.1) path P_h, the asymptotic distribution of $\widehat{\psi}$ has the following local uniformity property:

$$\sqrt{n}\big(\widehat{\psi}_n - \psi(P_{h(n)})\big) \overset{P_{h(n)}}{\rightsquigarrow} L_{P_0} \tag{4.12}$$

for any sequence $h(n) = O(1/\sqrt{n})$ and some probability measure L_{P_0} on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that does not depend on the path P_h. Local regularity requires the estimator to act uniformly on data from counterfactual distributions in a shrinking neighborhood of P_0 and rules out pathological Example 4.2. This condition links statistical properties of the estimator $\widehat{\psi}_n$ to analytic properties of the functional $\psi(P)$, and is used in Theorem 6 implicitly. Semiparametric efficiency theory uses (4.12) to derive a sharp characterization of the asymptotic distribution L_P and establish a bound on the asymptotic precision of regular estimators, see Hájek (1970) [38], van der Vaart (1991) [72]. Newey (1994) [56] uses this property to guarantee equality of the influence function $\widetilde{\psi}_P$ of the estimator and the Riesz representative $\nabla_{\text{F}}\psi_P$ of the functional. This paper uses local regularity to obtain counterfactual values of the functional from the asymptotic distribution of its estimators.

Before proving Theorem 6, we briefly unpack condition (4.12). Hodges' estimator is asymptotically linear but not locally regular. There are estimators that are regular but not asymptotically linear. The following more primitive conditions take advantage of the asymptotic linearity of an estimator and imply local regularity.

Fix a differentiable in quadratic mean path P_h and assume that estimator $\widehat{\psi}_n$ is asymptotically linear. To verify (4.12), it is sufficient that the linear representation (2.8) holds uniformly and continuously at P_0 on the path P_h:

Condition ALU. The remainder terms $o_{P(h),n}(1) = \sqrt{n}(\widehat{\psi}_n - \psi(P_h)) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \widetilde{\psi}_{P(h)}(X_i)$ of the estimator's asymptotic expansions based on random samples of size n from distributions P_h converge to zero as $n \to \infty$ in probability uniformly in h:

$$\lim_{n\to\infty} \sup_h P_h\big\{|o_{P^n(h),n}| > \epsilon\big\} = 0, \qquad \text{for every } \epsilon > 0. \tag{4.13}$$

Condition ALC. The variance function $h \mapsto \int_{\mathcal{X}} \widetilde{\psi}^2_{P(h)}\, dP_h$ of the asymptotic Gaussian distribution of the estimator is a continuous real-valued function of a real variable h.

These conditions strengthen asymptotic linearity by requiring that the sample averages of different influence functions have distributions that vary smoothly on smooth paths of data distributions, and approximate the distributions of the estimator uniformly well on the path P_h. The ALU condition can be verified with a uniform in P law of large numbers that extends the classical result by requiring uniform integrability, see Chung (1951) [22]. The ALC condition is straightforward to interpret.

Note that Hodges' estimator violates condition ALC because $P \mapsto P\widetilde{\psi}^2_{\text{H},P}$ has a jump discontinuity at every P with mean zero. DasGupta and Johnstone (2013) [23] show that the risk of $\widehat{\psi}_H$ is unbounded in a neighborhood around the threshold, which suggests that the remainder

terms in condition ALU may not be uniformly small near P with zero mean.

Lemma 17. Suppose the path $h \mapsto P_h$ is differentiable in quadratic mean (2.1), and suppose that estimator $\widehat{\psi}_n$ has the asymptotically linear representation (2.8) on P_h. If conditions ALU and ALC hold, then the estimator is locally regular on the path P_h.

Proof. Also see Bickel et al (1993) [13, Proposition 2.2.1, Appendix A.7]. We refer to Dudley (2002) [28] for details of probability theory and to Petrov (1975) [59] for results about uniform in P limit theorems.

Step 1: Conditions ALU and ALC imply that $h \mapsto \psi(P_h)$ is continuous. We first show that $\widehat{\psi}_n$ is uniformly consistent in P_h. Fix $\epsilon > 0$. From representation (2.8) and triangle inequality have

$$\big\{|\widehat{\psi}_n - \psi(P_h)| \le 2\epsilon\big\} \supset \big\{|\tfrac{1}{n}\textstyle\sum_{i=1}^n \widetilde{\psi}_{P(h)}(X_i)| \le \epsilon\big\} \cap \big\{|o_{P(h),n}(n^{-1/2})| \le \epsilon\big\}.$$

By De Morgan's law and subadditivity of measure, Markov's inequality it follows that

$$\begin{aligned} P_h\big\{|\widehat{\psi}_n - \psi(P_h)| > 2\epsilon\big\} &\le P_h\big\{|\tfrac{1}{n}\textstyle\sum_{i=1}^n \widetilde{\psi}_{P(h)}(x_i)| > \epsilon\big\} + P_h\big\{|o_{P(h),n}(n^{-1/2})| > \epsilon\big\}. \\ &\le \frac{1}{n\epsilon^2}\int_{\mathcal{X}} \widetilde{\psi}^2_{P(h)}\, dP_h + P_h\Big\{|o_{P(h),n}(n^{-1/2})| > \epsilon\Big\}. \end{aligned}$$

goes to zero as $n \to \infty$ uniformly in h under conditions ALU and ALC.

Next consider the space of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with the topology of convergence in distribution metrized by Prokhorov metric

$$d_{\mathrm{PR}}(L_1, L_2) = \inf\big\{\epsilon > 0\ ;\ L_1(A^\epsilon) \le L_2(A) + \epsilon \text{ all } A \in \mathcal{B}(\mathbb{R})\big\},$$

where $A^\epsilon := \{y \in \mathbb{R}\ ;\ |x - y| \text{ for some } x \text{ in } A\}$ is the ϵ-enlargement of set A. For point mass distributions, with δ_x denoting the point mass at $x \in \mathbb{R}$, convergence in law is equivalent to the convergence of the points of mass:

$$d_{\mathrm{PR}}(\delta_{\psi(P_h)}, \delta_{\psi(P_0)}) \to 0 \quad \text{if and only if} \quad \psi(P_h) \to \psi(P_0), \quad h \to 0.$$

Therefore it is sufficient to show that the Prokhorov distance above goes to zero.

By the triangle inequality, with $\widetilde{\psi}_* P_h^n$ denoting the push-forward measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by the estimator $\widetilde{\psi}_n$ of the random sample $(X_1, \ldots, X_n)$ under distribution P_h of the data,

$$d_{\mathrm{PR}}(\delta_{\psi(P_h)}, \delta_{\psi(P_0)}) \le d_{\mathrm{PR}}(\delta_{\psi(P_h)}, \widehat{\psi}_* P_h^n) + d_{\mathrm{PR}}(\widehat{\psi}_* P_h^n, \widehat{\psi}_* P_0^n) + d_{\mathrm{PR}}(\widehat{\psi}_* P_0^n, \delta_{\psi(P_0)}). \tag{4.14}$$

Let X, Y be any random variables in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and define the Ky Fan metric

$$d_{\mathrm{KF}}(X, Y) = \inf\big\{\epsilon > 0\ ;\ P[|X - Y| > \epsilon] \le \epsilon\big\},$$

which metrizes convergence in probability [see 28, Theorem 9.2.2], and can be used to define uniform consistency of estimators $\widehat{\psi}_n$ under different P_h. By [28, Theorem 11.3.5], the first term of the bound in (4.14) is further bounded by

$$d_{\mathrm{PR}}(\delta_{\psi(P_h)}, \widehat{\psi}_* P_h^n) \le d_{\mathrm{KF}}(\psi(P_h), \widehat{\psi}_n)$$

and is arbitrarily small uniformly in h as $n \to \infty$.

Because the path $h \mapsto P_h$ is differentiable in quadratic mean, it is also continuous in quadratic mean, which means that as $h \to 0$, $P_h \to P_0$ in total variation and therefore $P_h^n \to P_0^n$ in total variation [see 13, Appendix A.6]. It is immediate from definition of convergence in total variation, that the push-forward measures also converge $\widehat{\psi}_* P_h^n \to \widehat{\psi}_* P_0^n$ in total variation as $h \to 0$. But this implies convergence in distribution $d_{\mathrm{PR}}(\widehat{\psi}_* P_h^n, \widehat{\psi}_* P_0^n) \to 0$ as $h \to 0$, as can be seen by the Portmanteau characterization of convergence in distribution [see 71, Lemma 2.2]. Conclude that the bound in eq. (4.14) can be made arbitrarily small by first choosing n sufficiently large

to control the first and the third terms (uniformly in h), and the choosing h sufficiently close to 0 to control the second term.

Fix a sequence $h(n) = O(1/\sqrt{n})$. Let $P_k = P(k) = P_{h(k)}$.

Step 2: The push-forward measure of the influence functions $\widetilde{\psi}_{P(k)_*}P_k \rightsquigarrow \widetilde{\psi}_{P(0)_*}P_0$ converge in distribution as $k \to \infty$. From differentiability in quadratic mean of the path $h \mapsto P_h$ at P_0, it follows that product measures P_n^n and P_0^n are mutually contiguous, by local asymptotic normality and Le Cam's third lemma [see e.g. 71, Theorem 7.2 and Example 6.7]. Therefore $P_n^n(|o_{P(0),n}(1)| > \epsilon) \to 0$ for every $\epsilon > 0$ by the contiguity $P_{h(n)}^n \lhd P_0^n$ and linear representation eq. (2.8) at P_0. Also $P_n^n(|o_{P(n),n}(1)| > \epsilon) \to 0$ for every $\epsilon > 0$ by condition ALU. Thus,

$$\begin{aligned} o_{P(0),n}(1) - o_{P(n),n}(1) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\widetilde{\psi}_{P(n)}(X_i) + \psi(P_n) - \widetilde{\psi}_P(X_i) - \psi(P) \right] \\ &= o_{P(n),n}(1) \\ &= o_{P(0),n}(1) \end{aligned}$$

the difference in the linear representations of the estimator at P_n and P_0 vanishes under both P_n^n and P_0^n by mutual contiguity of these sequences. If follows that the terms of the series above vanish in probability

$$\left[\widetilde{\psi}_{P(n)}(X_i) + \psi(P_n) - \widetilde{\psi}_P(X_i) - \psi(P) \right] = o_{P(0)}(1) = o_{P(n)}(1).$$

Therefore, the push-forward probability measures converge in distribution:

$$\left(\widetilde{\psi}_{P(n)} + \psi(P_n) \right)_* P_0 \rightsquigarrow \left(\widetilde{\psi}_{P(0)} + \psi(P_0) \right)_* P_0.$$

Since by Step 1, $\delta_{\psi(P_n)} \rightsquigarrow \delta_{\psi(P_0)}$, by the triangle inequality of e.g. Prokhorov metric defined above, by $P_n \to P_0$ in total variation, we must have

$$\left(\widetilde{\psi}_{P(n)} \right)_* P_n \rightsquigarrow \left(\widetilde{\psi}_{P(0)} \right)_* P_0. \tag{4.15}$$

Step 3: Convergence eq. (4.15) and continuity of variance $\int_{\mathcal{X}} \widetilde{\psi}^2_{P(k)} \, P_k \to \int_{\mathcal{X}} \widetilde{\psi}^2_{P(0)} \, P_0$ imply the Lindeberg condition

$$\lim_{\lambda \to \infty} \sup_k \mathrm{E}_{P_k} \left[\widetilde{\psi}^2_{P(k)} \mathbb{1}_{\{|\widetilde{\psi}^2_{P(k)}| > \lambda\}} \right] = 0.$$

This follows by e.g. [71, Theorem 2.20]. By e.g. Petrov (1975) [59, Theorem V.3.8], the triangular array

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\psi}_{P_h(n)}(X_{i,n}), \quad X_{1,n}, \ldots, X_{n,n} \sim P_{h(n)}^n$$

admits the Central Limit Theorem $S_n \rightsquigarrow \mathcal{N}\left(0, \int_{\mathcal{X}} \widetilde{\psi}^2_{P_0} \, dP_0\right)$. But this means that $\widehat{\psi}_n$ is regular on P_h at P_0. Similarly conclude regularity for any other point on the curve P_h. □

Proof of Theorem 6. We follow [13, Theorem 3.3.1] closely in the first part of the proof. Fix a differentiable in quadratic mean path P_h and let $v \in L_0^2(P_0)$ denote its score function. Fix a sequence $h(n) = t_n/\sqrt{n}$, with $t_n \to t$ Let $P_n = P(n) = P_{h(n)}$. By the asymptotic linearity eq. (2.8) of $\widetilde{\psi}_n$, and local asymptotic normality of the parametric model P_h [see e.g. 71, Theorem 7.2],

$$\sqrt{n} \begin{bmatrix} \widehat{\psi}_n - \psi(P_0) \\ \log \prod_{i=1}^n \frac{dP_{h(n)}}{dP_0}(X_i) \end{bmatrix} \overset{P_0}{\rightsquigarrow} \mathcal{N}\left(\begin{bmatrix} 0 \\ -\Sigma_{22}/2 \end{bmatrix}, \Sigma \right),$$

where $\Sigma = [\Sigma_{ij}]_{ij=1,2}$, $\Sigma_{11} = \mathrm{E}_{P_0}\left[\widetilde{\psi}^2_{P_0}\right]$, $\Sigma_{12} = \mathrm{E}_{P_0}\left[\widetilde{\psi}_{P_0}(tv)\right]$, $\Sigma_{22} = \mathrm{E}_{P_0}\left[(tv)^2\right]$.

By Le Cam's third lemma [see 71, Example 6.5], the distribution of the estimator under the local alternatives follows by contiguity:

$$\sqrt{n}(\widehat{\psi}_n - \psi(P_0)) \overset{P_n}{\rightsquigarrow} \mathcal{N}(\Sigma_{12}, \Sigma_{11}). \tag{4.16}$$

By the regularity of estimator sequence $\widehat{\psi}_n$, also have the limit in distribution under local alternatives:

$$\sqrt{n}(\widehat{\psi}_n - \psi(P_n)) \overset{P_n}{\rightsquigarrow} \mathcal{N}(0, \Sigma_{11}). \tag{4.17}$$

Combining (4.16) and (4.17), we conclude that

$$\sqrt{n}(\psi(P_n) - \psi(P_0)) \to \Sigma_{12} = \int_{\mathcal{X}} \widetilde{\psi}_{P_0}(tv)\, dP_0.$$

This can be written as

$$\frac{\psi(P_{0+t_n/\sqrt{n}}) - \psi(P_0)}{t_n\sqrt{n}} = \int_{\mathcal{X}} \widetilde{\psi}_{P_0} v\, dP_0 + o(1). \tag{4.18}$$

Next, we follow the proof of van der Vaart (1991) [72, Theorem 2.1]. For any sequence of real numbers $r_m \to 0$, let n_m be the subsequence such that

$$(1 + n_m)^{-1/2} < r_m \leq n_m^{-1/2}.$$

Let $\widetilde{h}(n) = r(m)$ for $n = n_m$ and let $\widetilde{h}(n) = n^{-1/2}$ for $n \notin \{n_1, n_2, \ldots\}$. By construction $\widetilde{h}(n) = O(1/\sqrt{n})$ so that the limit (4.18) holds for the sequence $\widetilde{h}(n)$ and also for its subsequence $r(m)$. Conclude that the derivative of ψ along the path P_h exists and is given by the follow representation:

$$\tfrac{d}{dh}_{|h=0}\psi(P_h) = \int_{\mathcal{X}} \widetilde{\psi}_{P_0} v\, dP_0.$$

Since the integral mapping of the tangent set $\mathcal{T}_P$ is linear and bounded, conclude that ψ is pathwise differentiable at P_0 and that $\widetilde{\psi}_{P_0}$ is its information gradient score. The hardest submodel $\{P_{\text{F},\nu,h,P_0}\}_{h\in J}$ exists by Theorem 13. □

This theorem follows the result of van der Vaart (1991) in that it has differentiability of the target and control parameters as a conclusion. We want to emphasize that the purpose of the theorem is very different from the results of Bickel, Klaasen, Ritov, and Wellner (1993), van der Vaart (1991), Newey (1994) and Ichimura and Newey (2015) and other important contributions to semiparametric efficiency theory and asymptotic inference in econometrics. Here we use the regularity of the functional and the estimator to find a rationalizable counterfactual distribution P_h and to calculate the counterfactual value of the parameter and the estimator at that distribution. By contrast, semiparametric efficiency uses regularity to establish a lower bound on the asymptotic precision of estimators, whereas Newey uses regularity to find the asymptotic distribution of complex estimators.

The result relies on the key assumption that estimators $\widehat{\psi}, \widehat{\nu}$ are regular. Most classical estimators employed by economists are regular, however, estimators that rely on model selection or reduce variance with shrinkage can be nonregular. Estimators that have a high-dimensional nuisance component and use modern machine learning tools may have nonregular behavior, as pointed out by Belloni, Chernozhukov and Hansen (2013) [11]. For nonregular estimators, analytic counterfactuals developed here do not apply. Our policy analysis highlights the importance of the work by Newey on techniques to *prove* that a complex estimator is regular, and by Chernozhukov with many coauthors on setting up estimators in the high-dimensional setting that *are* regular.

References

[1] H. Amann. *Ordinary differential equations: an introduction to nonlinear analysis*. Vol. 13. Walter de Gruyter, 1990.

[2] S.-i. Amari. *Differential-geometrical methods in statistics*. Vol. 28. Springer-Verlag, 1985.

[3] S.-i. Amari and H. Nagaoka. *Methods of Information Geometry*. Translations of mathematical monographs. American Mathematical Society, 2000. ISBN: 9780821843024.

[4] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[5] L. Ambrosio and G. Savaré. "Gradient flows of probability measures". *Handbook of differential equations: evolutionary equations* 3 (2007), pp. 1–136.

[6] I. Andrews, M. Gentzkow, and J. M. Shapiro. "Measuring the Sensitivity of Parameter Estimates to Estimation Moments". *The Quarterly Journal of Economics* (2017).

[7] I. Andrews, M. Gentzkow, and J. M. Shapiro. *On the Informativeness of Descriptive Statistics for Structural Estimates*. Working Paper. 2018.

[8] T. Armstrong and M. Kolesár. "Sensitivity analysis using approximate moment condition models" (2018).

[9] O. E. Barndorff-Nielsen, D. R. Cox, and N. Reid. "The role of differential geometry in statistical theory". *International statistical review* 54.1 (1986), pp. 83–96.

[10] J. M. Begun et al. "Information and Asymptotic Efficiency in Parametric-Nonparametric Models". *Ann. Statist.* 11.2 (June 1983), pp. 432–452.

[11] A. Belloni, V. Chernozhukov, and C. Hansen. "Inference on treatment effects after selection among high-dimensional controls". *The Review of Economic Studies* 81.2 (2014), pp. 608–650.

[12] J.-D. Benamou and Y. Brenier. "A computational fluid mechanics solution to the Monge - Kantorovich mass transfer problem". *Numerische Mathematik* 84.3 (2000), pp. 375–393.

[13] P. J. Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[14] A. S. Blinder. "Wage discrimination: reduced form and structural estimates". *Journal of Human resources* (1973), pp. 436–455.

[15] S. Bonhomme and M. Weidner. "Minimizing sensitivity to model misspecification". *arXiv preprint arXiv:1807.02161* (2018).

[16] G. Burdet, P. Combe, and H. Nencka. "On real Hilbertian info-manifolds". *AIP Conference Proceedings*. Vol. 553. 1. AIP. 2001, pp. 153–158.

[17] M. P. d. Carmo. *Differential geometry of curves and surfaces*. Prentice-Hall, 1976.

[18] M. P. d. Carmo. *Riemannian geometry*. Birkhäuser, 1992.

[19] A. Cena and G. Pistone. "Exponential statistical manifold". *Annals of the Institute of Statistical Mathematics* 59.1 (2007), pp. 27–56.

[20] V. Chernozhukov, I. Fernández-Val, and B. Melly. "Inference on counterfactual distributions". *Econometrica* 81.6 (2013), pp. 2205–2268.

[21] T. Christensen and B. Connault. "Counterfactual Sensitivity and Robustness". *arXiv preprint arXiv:1904.00989* (2019).

[22] K. L. Chung. "The Strong Law of Large Numbers". *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1951, pp. 341–352.

[23] A. DasGupta and I. M. Johnstone. *Risk and Bayes Risk of Thresholding and Superefficient Estimates, Regular Variation, and Optimal Thresholding*. 2013.

[24] P. J. Dhrymes. *Mathematics for econometrics*. Tech. rep. Springer, 1978.

[25] J. DiNardo, N. M. Fortin, and T. Lemieux. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach". *Econometrica* 64.5 (1996), pp. 1001–1044.

[26] D. L. Donoho and R. C. Liu. "Geometrizing rates of convergence, II". *The Annals of Statistics* (1991), pp. 633–667.

[27] D. L. Donoho and R. C. Liu. "Geometrizing rates of convergence, III". *The Annals of Statistics* (1991), pp. 668–701.

[28] R. Dudley. *Real Analysis and Probability*. Vol. 74. Cambridge University Press, 2002.

[29] B. Efron. "Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency)". *Ann. Statist.* 3.6 (Nov. 1975), pp. 1189–1242.

[30] S. Firpo, N. M. Fortin, and T. Lemieux. "Unconditional quantile regressions". *Econometrica* 77.3 (2009), pp. 953–973.

[31] S. Firpo, N. Fortin, and T. Lemieux. "Decomposing wage distributions using recentered influence function regressions". *Econometrics* 6.2 (2018), p. 28.

[32] G. B. Folland. *Advanced calculus*. Prentice Hall, 2002.

[33] N. Fortin, T. Lemieux, and S. Firpo. "Decomposition methods in economics". *Handbook of labor economics*. Vol. 4. Elsevier, 2011, pp. 1–102.

[34] K. Fukumizu. "Exponential manifold by reproducing kernel hilbert spaces". *Algebraic and Geometric mothods in statistics* (2009), pp. 291–306.

[35] W. Gangbo and R. J. McCann. "The geometry of optimal transportation". *Acta Mathematica* 177.2 (1996), pp. 113–161.

[36] M. Gentzkow and J. M. Shapiro. *Measuring the Sensitivity of Parameter Estimates to Sample Statistics*. Working Paper. 2015.

[37] M. Grasselli. "Dual connections in nonparametric classical information geometry". *Annals of the Institute of Statistical Mathematics* 62.5 (2010), pp. 873–896.

[38] J. Hájek. "A characterization of limiting distributions of regular estimates". *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 14.4 (1970), pp. 323–330.

[39] A. R. Hall and A. Inoue. "The large sample behaviour of the generalized method of moments estimator in misspecified models". *Journal of Econometrics* 114.2 (2003), pp. 361–394.

[40] F. R. Hampel. "The influence curve and its role in robust estimation". *Journal of the american statistical association* 69.346 (1974), pp. 383–393.

[41] J. A. Hausman and W. K. Newey. "Nonparametric estimation of exact consumers surplus and deadweight loss". *Econometrica: Journal of the Econometric Society* (1995), pp. 1445–1476.

[42] J. J. Heckman and E. J. Vytlacil. "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation". *Handbook of econometrics* 6 (2007), pp. 4779–4874.

[43] H. Hotelling. "Spaces of statistical parameters". *Bull. Amer. Math. Soc.* 36.3 (Mar. 1930), p. 191.

[44] P. J. Huber. *Robust statistics*. Wiley, 1981.

[45] H. Ichimura and W. K. Newey. "The influence function of semiparametric estimators". *arXiv preprint arXiv:1508.01378* (2017).

[46] G. W. Imbens. "One-step estimators for over-identified generalized method of moments models". *The Review of Economic Studies* 64.3 (1997), pp. 359–383.

[47] Y. A. Koshevnik and B. Y. Levit. "On a Non-Parametric Analogue of the Information Matrix". *Theory of Probability & Its Applications* 21.4 (1976), pp. 738–753.

[48] S. Lang. *Fundamentals of Differential Geometry*. Graduate Texts in Mathematics. Springer-Verlag New York, 1999. ISBN: 978-0-387-98593-0.

[49] L. LeCam and L. Schwartz. "A necessary and sufficient condition for the existence of consistent estimates". *The Annals of Mathematical Statistics* 31.1 (1960), pp. 140–150.

[50] B. Y. Levit. "On optimality of some statistical estimates". *Proceedings of the Prague symposium on asymptotic statistics*. Vol. 2. 1974, pp. 215–238.

[51] B. Y. Levit. "On the Efficiency of a Class of Non-Parametric Estimates". *Theory of Probability & Its Applications* 20.4 (1975), pp. 723–740.

[52] P. C. Mahalanobis. "On the generalized distance in statistics". National Institute of Science of India. 1936.

[53] R. von Mises. "On the asymptotic distribution of differentiable statistical functions". *The annals of mathematical statistics* 18.3 (1947), pp. 309–348.

[54] J. Neveu. *Mathematical foundations of the calculus of probability*. Holden-day, 1965.

[55] W. K. Newey. "Semiparametric efficiency bounds". *Journal of applied econometrics* 5.2 (1990), pp. 99–135.

[56] W. K. Newey. "The asymptotic variance of semiparametric estimators". *Econometrica: Journal of the Econometric Society* (1994), pp. 1349–1382.

[57] N. J. Newton. "An infinite-dimensional statistical manifold modelled on Hilbert space". *Journal of Functional Analysis* 263.6 (2012), pp. 1661–1681.

[58] R. Oaxaca. "Male-female wage differentials in urban labor markets". *International economic review* (1973), pp. 693–709.

[59] V. V. Petrov. *Sums of independent random variables*. Vol. 82. Springer, 1975.

[60] G. Pistone. "Nonparametric information geometry". *International Conference on Geometric Science of Information*. Springer. 2013, pp. 5–36.

[61] G. Pistone and C. Sempi. "An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one". *The Annals of Statistics* 23.5 (1995), pp. 1543–1561.

[62] C. R. Rao. "Information and the Accuracy Attainable in the Estimation of Statistical Parameters". *Bulletin of Cal. Math. Soc.* 37.3 (1945), pp. 81–91.

[63] J. Robins et al. "Higher order influence functions and minimax estimation of nonlinear functionals". *Probability and statistics: essays in honor of David A. Freedman*. Institute of Mathematical Statistics, 2008, pp. 335–421.

[64] H. Royden and P. Fitzpatrick. "Real analysis (4th Edtion)". *New Jersey: Printice-Hall Inc* (2010).

[65] T. A. Severini, G. Tripathi, et al. "Semiparametric efficiency bounds for microeconometric models: A survey". *Foundations and Trends® in Econometrics* 6.3–4 (2013), pp. 163–397.

[66] C. Stein. "Efficient Nonparametric Testing and Estimation". *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1956, pp. 187–195.

[67] E. M. Stein and R. Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

[68] J. H. Stock. "Nonparametric policy analysis". *Journal of the American Statistical Association* 84.406 (1989), pp. 567–575.

[69] J. H. Stock. "Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits". *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (1991), pp. 77–98.

[70] T. M. Stoker. "Consistent estimation of scaled coefficients". *Econometrica: Journal of the Econometric Society* (1986), pp. 1461–1481.

[71] A. W. van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 1998.

[72] A. W. van der Vaart. "On differentiable functionals". *The Annals of Statistics* (1991), pp. 178–204.

[73] A. Vanhems. "Nonparametric study of solutions of differential equations". *Econometric Theory* 22.1 (2006), pp. 127–157.

[74] H. White. "Using least squares to approximate unknown regression functions". *International Economic Review* (1980), pp. 149–170.