

Machine Learning Inference on Inequality of Opportunity*

Juan Carlos Escanciano Joël R. Terschuur

October 4, 2023

Abstract

Equality of opportunity has emerged as an important ideal of distributive justice. Empirically, Inequality of Opportunity (IOp) is measured in two steps: first, an outcome (e.g., income) is predicted given individual circumstances; and second, an inequality index (e.g., Gini) of the predictions is computed. Machine Learning (ML) methods are tremendously useful in the first step. However, they can cause sizable biases in IOp since the bias-variance trade-off allows the bias to creep in the second step. We propose a simple debiased IOp estimator robust to such ML biases and provide the first valid inferential theory for IOp. We demonstrate improved performance in simulations and report the first unbiased measures of income IOp in Europe. Mother's education and father's occupation are the circumstances that explain the most. Plug-in estimators are very sensitive to the ML algorithm, while debiased IOp estimators are robust. These results are extended to a general U-statistics setting.

JEL Classification: C13; C14; C21; D31; D63

Keywords: local robustness, orthogonal moments, machine learning, U-statistics, Inequality of Opportunity.

R package: <https://joelters.github.io/home/code/>

***Escanciano** (corresponding author): Universidad Carlos III de Madrid (jescanci@eco.uc3m.es). **Terschuur**: Universidad Carlos III de Madrid (jrobert@eco.uc3m.es). Postal address Calle Madrid 126, Edificio 15, Getafe, Madrid 28903. Research funded by Ministerio de Ciencia e Innovación, grant ECO2017-86675-P, MCI/AEI/FEDER/UE, grant PGC 2018-096732-B-100, grant PID2021-127794NB-I00, grant PRE2020-092794 and Comunidad de Madrid, grants EPUC3M11 (VPRICIT) and H2019/HUM-589. We thank conference participants at the IAAE 2022, Tenth ECINEQ meeting 2023 and AiE conference in honor of Joon Park.

1 Introduction

The moral standing of inequality has been part of the social debate for centuries. While unequal returns might foster effort and innovation, an uneven distribution of life outcomes such as income, wealth, education or health, can disrupt society and lead to unrest and political polarization. A concept that has gained a broad consensus is the one of Inequality of Opportunity (IOp). Inequalities which stem from circumstances outside the control of the individual (henceforth just referred to as circumstances) are generally deemed unfair. Examples of such circumstances are parental education, biological sex, color of the skin or social origin. It is precisely the unchosen nature of circumstances that fuels a debate on gender gaps, racial discrimination or the importance of parental resources for many vital outcomes. Not only is IOp considered unfair, but it might also hinder economic growth (e.g. [Marrero and Rodríguez \(2013\)](#), [Ferreira et al. \(2018\)](#), [Aiyar and Ebeke \(2020\)](#) or [Carranza \(2020\)](#)). While inequalities based on different returns to different levels of effort can spur innovation and growth, inequalities based on circumstances might entail an important waste of talents and capabilities. If the determinants of success are circumstances, we lose the contribution of an important share of society. Moreover, perceptions about IOp shape redistributive preferences and hence the design of social policy (e.g. [Alesina and La Ferrara \(2005\)](#)).

The well-grounded theoretical framework developed in the seminal contributions by [Van De Gaer \(1993\)](#), [Fleurbaey \(1995\)](#) and [Roemer \(1998\)](#) has fueled a growing empirical literature quantifying IOp, see the reviews in [Roemer and Trannoy \(2016\)](#), [Ramos and Van de Gaer \(2016\)](#) or [Ferreira and Peragine \(2016\)](#). However, the state-of-the-art empirical tools for IOp are not satisfactory: current methods lead to highly biased estimates and no valid inferential methods are available. By expanding novel results at the intersection of semiparametric econometrics and machine learning literatures we provide the first valid and robust inferential methods for IOp.

The leading measure of IOp is the Gini coefficient of fitted values or predictions, possibly obtained from a Machine Learning (ML) first step. This estimator is referred to as the plug-in estimation in this paper. The main idea is that if one can predict an individual outcome using circumstances, then there is IOp. By predicting outcomes with circumstances and then measuring the inequality in the distribution of predictions we can quantify IOp. For example, suppose that the income of any individual was the same as the income of his/her parents. Then, we would perfectly predict income given parental income so all income inequality would be due to circumstances. Of course, perfect prediction will not be possible in practice, and the prediction step will introduce an additional layer of uncertainty into the measurement of IOp. The two-step nature of measures of IOp creates important challenges for the development of inferential methods, as we explain below.

The recent and increasing use of ML in IOp addresses an important and long-standing problem in the literature: how to efficiently exploit the information of all observed circumstances in the predictions avoiding ad-hoc partitions of circumstances into types. Traditional estimation methods have made it either prohibitively costly to include all observable circumstances (e.g. fully non-parametric methods) or have imposed far too much structure on the relationship between the outcome and the circumstances (e.g. log-linear low-dimensional regression). ML techniques, which can handle high dimensional problems, can overcome this problem by trading-off variance and bias in the predictions. Machine learners such as random forests, boosting, lasso or neural networks, among many others, can be used to obtain high-quality predictions of the outcome. This has motivated the recent use of ML for IOp, such as Conditional Forest (see Brunori et al. (2021), Brunori et al. (2019a), Brunori and Neidhöfer (2021), Salas-Rojo and Rodríguez (2022) or Carranza (2022)) or Lasso in Hufe et al. (2022c). The increasing availability of high-quality datasets with a fine degree of information on circumstances will make the applications of ML methods even more attractive.

The first contribution of this paper is to show that the high-quality predictions of ML lead to sizable regularization and model selection biases in the measurement of IOp. Intuitively, machine learners trade off bias and variance; meaning that some bias is allowed if it improves the prediction. This bias is not a problem for prediction, but it is for inference on IOp if the bias creeps into the second step. Empirically, the bias is so large that makes standard inference procedures invalid for commonly observed sample sizes, see Figure 1 in Section 2.2 for illustration.

The second contribution is the proposal of a novel and extremely simple debiased IOp estimator. To introduce the IOp estimator, let Y_i be an outcome of interest (e.g., income, education or health) and let $\hat{\gamma}(X_i)$ be the prediction at the vector of circumstances X_i , e.g. from a Random Forest fit. The new debiased estimator based on a sample $\{Y_i, X_i\}_{i=1}^n$ is

$$\frac{\sum_{i<j} \text{sgn}(\hat{\gamma}(X_i) - \hat{\gamma}(X_j))(Y_i - Y_j)}{\sum_{i<j} Y_i + Y_j}, \quad (1.1)$$

where $\sum_{i<j} = \sum_{i=1}^{n-1} \sum_{j=i+1}^n$ and $\text{sgn}(x)$ is the sign function, which equals 1, 0 and -1 , for a positive, zero and negative number x , respectively. The estimator is extremely simple to implement.¹ It is like the standard Gini coefficient but with the sign of outcome differences replaced by the fitted values differences.² By way of comparison, the plug-in IOp estimator is like (1.1) but with the outcome variables replaced by predictions. The new estimator in (1.1) leads to valid and robust inference for a great variety of ML first steps. In contrast to the vast

¹See Terschuur (2022) (or <https://joelsters.github.io/home/code/>) for an R package which implements the debiased IOp estimator and the related inferences.

²The standard Gini coefficient is $\sum_{i<j} |Y_i - Y_j| / \sum_{i<j} (Y_i + Y_j) = \sum_{i<j} \text{sgn}(Y_i - Y_j)(Y_i - Y_j) / \sum_{i<j} (Y_i + Y_j)$.

majority of debiased methods using ML, no additional nuisance parameters are necessary to obtain (1.1). Surprisingly, the estimator coincides with the estimator of the concentration index of Y_i with respect to $\gamma(X_i)$ yielding two interesting results (see [Yitzhaki and Olkin \(1991\)](#) for more information on concentration indices and concentration curves): (i) IOp can be interpreted as a concentration index, (ii) concentration indices of Y_i with respect to some estimated quantities are locally robust in a sense defined in this paper. Independently of our paper, [Heuchenne and Jacquemain \(2022\)](#) provide a bootstrap-based inference for IOp in the special case of a single index model with a strictly increasing link function and also arrive to a concentration index. Our estimator can be seen as a nonparametric version of their Lorenz regression allowing for ML first steps. For high dimensional settings or ML first steps, a cross-fitting version of the estimator in (1.1) is proposed below in (2.6).

We obtain these results as an application of our more general theory that covers a large class of semiparametric U-Statistics, i.e., U-statistics with high dimensional estimated parameters. Additional example applications include the analysis of variance in high dimensions ([Lou et al. \(2023\)](#)), measures of economic polarization ([Duclos et al. \(2004\)](#)), optimal risk in the bipartite ranking problem ([Cléménçon et al. \(2008\)](#) or [Werner \(2021\)](#)), distance-based estimators in semiparametric conditional moment restrictions ([Domínguez and Lobato \(2004\)](#)), semiparametric estimators of binary models with endogeneity ([Blundell and Powell \(2004\)](#)), or quadratic functionals of the marginal distributions of potential outcomes in treatment effects ([Wu et al. \(2014\)](#) or [Mao \(2018\)](#)), among many others. We give a general construction of orthogonal quadratic moment functions which can be used to obtain debiased estimators and valid inferences in all these settings.

Our results on U-statistics generalize the results of the recent and growing debiasing literature, see, e.g., [Chernozhukov et al. \(2022\)](#). We innovate in constructing *quadratic* orthogonal moments for semiparametric U-statistics, rather than orthogonal moments for GMM. This innovation entails new conceptual and non-trivial methodological problems. For example, unlike GMM moments, quadratic orthogonal moments are not unique for a given identifying moment and first step limit, but they all share a unique Hajek projection, and therefore, a unique first order asymptotics, as we show below. [Robins et al. \(2017\)](#) and [Rajarshi Mukherjee \(2017\)](#) use U-statistics to characterize higher order influence functions for second and higher order effects. Our interest in U-statistics comes from the application to IOp and the simpler first-order effects. Our work is also related to [Chiang et al. \(2021a\)](#) and [Chiang et al. \(2021b\)](#), who proposed to use orthogonal moments for multiway clustered sampling and dyadic regressions, respectively. We derive orthogonal moments in this paper.

The orthogonality of the proposed moment functions is such that the estimation of first steps has no local effect on the parameters of interest, a property also referred to as local robustness in the literature. Semiparametric U-statistics have been treated in [Powell \(1987\)](#),

Sherman (1994) and Honoré and Powell (2005) and asymptotic theory has been derived using uniform versions of the Hoeffding decomposition under Donsker conditions (see e.g. Nolan and Pollard (1988) or Arcones and Giné (1993)). Unfortunately, Donsker conditions do not hold or are not known to hold in high dimensional settings. In this paper, we propose the use of orthogonal moments and cross-fitting to avoid Donsker conditions, as in Chernozhukov et al. (2018) but for U-statistics. Our asymptotic theory provides an alternative set of mild conditions to the more traditional methods cited in this paragraph.

The application of our general results to IOp faces some challenges and uncovers some interesting and surprising findings. First, the standard measure of IOp is not locally robust, which makes standard inference for Gini coefficients invalid and plug-in measures highly biased with ML first steps. To obtain these results, we address an additional technical challenge, which is a lack of differentiability of the identifying moment function for IOp. Second, from the numerous possible correction terms and orthogonal moments, we identify one that leads to a simple debiased IOp estimator. Surprisingly, and unlike in most applications of the existing debiased literature, the simple debiased IOp estimator does not require estimating additional nuisance parameters, making its practical implementation straightforward.

The performance of our IOp estimator is evaluated through Monte Carlo simulations using different machine learners for the first step estimation. We employ regularized linear regression techniques such as Lasso and Ridge and tree-based ensemble methods such as Random Forests (RF), Conditional Inference Forests (CIF), Extreme Gradient Boosting (XGB) and Catboosting (CB). Lasso, Ridge and RF are well-known machine learners in econometrics. CIF were developed as an alternative to RF in Hothorn et al. (2006) and they are used by the IOp literature, for instance in Brunori and Neidhöfer (2021), Brunori et al. (2021), Salas-Rojo and Rodríguez (2022) or Carranza (2022). Boosting methods, such as XGB and CB, are popular in machine learning competitions (see Chen and Guestrin (2016) and Prokhorenkova et al. (2018)). Our simulations confirm the high biases for inference on IOp from using ML first steps, particularly for the method recommended in the applied literature (CIF). Our debiased estimator corrects this bias, it is far less sensitive to the ML first step than the commonly used plug-in method and delivers valid inferences for the full battery of machine learners considered.

The empirical contribution of this paper is to use the first debiased IOp estimations of income IOp in 29 European countries using the 2019 cross-sectional European Union Statistics on Income and Living Conditions (EU-SILC) survey. This survey is one of the main references for the analysis of the income distribution and poverty in Europe. In the years 2005, 2011 and 2019 it includes a module on intergenerational transmission of disadvantages with information on circumstances. We restrict our attention to the year 2019 since it is the most recent one and contains the richest set of circumstances. The 2019 wave has been

used in Carranza (2022) to measure income IOp in Europe with plug-in Mean Logarithmic Deviation based measures and in Terschuur (2022). Hufe et al. (2022b) also uses the 2019 wave to assess multidimensional IOp using gender, parental education and parental occupation as circumstances. Brunori et al. (2019b) and Brunori et al. (2021) use cross-validation for model selection in the first step and CIF to measure IOp for EU-SILC 2011. Marrero and Rodríguez (2012) use EU-SILC 2005 to estimate IOp in Europe. Our paper is the first one to estimate IOp in EU-SILC 2019 using a variety of machine learners and debiased Gini IOp estimators with theoretically valid inference guarantees.

The countries with the highest IOp are Romania and Bulgaria where around 60% of income inequality can be explained by circumstances, while Denmark is the country with the lowest IOp. Nordic countries, Germany, the Netherlands and some Eastern countries have low IOp. Southern countries have high IOp with Greece and Italy having 39% and Spain having 43% of total inequality explained by circumstances. We find that the plug-in estimator tends to overestimate IOp in general and particularly so for Denmark. The difference between our debiased estimates and the biased plug-in estimates of relative IOp (i.e. fraction of inequality explained by circumstances), can be as large as 25 percentage points in the case of Denmark. In the rest of the countries the difference is smaller but we still observe differences of 5-10 percentage points. Hence, plug-in estimators give a biased view of IOp in Europe.

Our debiased IOp estimates are also much less sensitive to the choice of the machine learner than commonly used plug-in measures. We estimate relative IOp with six different machine learners with both the debiased and the plug-in estimator and see that plug-in estimates using different machine learners are much more dispersed than the debiased estimates. Plug-in estimates for the same country but using different machine learners can be up to 60 percentage points apart. Even differences of 30-40 percentage points are not uncommon. In contrast, debiased estimates are much more concentrated and do not exhibit such sizeable variations across machine learners. While hard to compare with previous empirical results due to the use of plug-in methods and older data, the traditional divide of North and South into high and low IOp remains in our results.

We also explore which circumstances contribute the most towards explaining inequality. We find that parental education and occupation are the most important circumstances explaining unfair inequality in European countries. Our results complement the empirical findings in Hufe et al. (2022a), who also find these circumstances to be the most important drivers of the increasing unfair inequality in the US during the period 1980-2014. However, we are able to observe differences between the impact of maternal and paternal characteristics. While in the case of the mothers it is the education that matters most, in the case of the fathers the occupation is the most important variable.

A crucial improvement of our methodology is the ability to perform inference. This allows to perform statistical tests of IOp between independent populations, across time, or between different sets of circumstances. The ability to do inference is paramount to evaluate changes in IOp before and after a policy has been implemented or across time. Hence, we believe the inferential tools provided here can be of great use for policy design aiming at reducing IOp.

The rest of the paper is organized as follows. Section 2 introduces the IOp setting, some general concepts and the IOp estimator. Section 3 covers how to construct orthogonal moment functions generally. Section 4 shows the performance of the debiased IOp estimator in Monte Carlo simulations. Finally, Section 5 studies IOp in Europe. Mathematical proofs are reported in an Appendix. The Online Appendix contains complementary results.

2 Methodology

We have independent and identically distributed (i.i.d.) data $W_i = (Y_i, X_i)$, for $i = 1, \dots, n$, distributed with unknown distribution F_0 , an unknown first step function γ_0 and a finite dimensional parameter of interest θ in some set $\Theta \subseteq \mathbb{R}^k$. We can form pairs (W_i, W_j) , where W_j is an independent copy of W_i , with realizations (w_i, w_j) . $\mathbb{E}[\cdot]$ is the expectation under F_0 , $\gamma(F)$ is the plim of a first step estimator $\hat{\gamma}$ under F , like in Newey (1994), and $\gamma_0 = \gamma(F_0)$.

2.1 Inequality of Opportunity

The leading measure of IOp is given by the Gini coefficient of fitted values

$$\theta_0 = \frac{\mathbb{E}[|\gamma_0(X_i) - \gamma_0(X_j)|]}{\mathbb{E}[\gamma_0(X_i) + \gamma_0(X_j)]}.$$

Here X_i is a vector of circumstances individual i did not choose, such as parental wealth/income, parental education, gender, color of the skin or social origin, but many other variables including high dimensional genetic information or even infinite dimensional (i.e. functional) circumstances, as in Chang et al. (2023), might be available. The function

$$g(w_i, w_j, \gamma, \theta) = (\gamma(x_i) + \gamma(x_j))\theta - |\gamma(x_i) - \gamma(x_j)|,$$

is an identifying moment function in the sense that

$$\mathbb{E}[g(W_i, W_j, \gamma_0, \theta)] = 0 \iff \theta = \theta_0.$$

A natural and popular estimator of θ_0 is the sample Gini coefficient of fitted values

$$\hat{\theta}^P = \frac{\sum_{i < j} |\hat{\gamma}(X_i) - \hat{\gamma}(X_j)|}{\sum_{i < j} (\hat{\gamma}(X_i) + \hat{\gamma}(X_j))},$$

for a first step $\hat{\gamma}$ such as CIF, see Brunori et al. (2021), Brunori et al. (2019a) or Brunori and Neidhöfer (2021). The plug-in estimator $\hat{\theta}^P$ solves the sample analog of the quadratic moment $\mathbb{E}[g(W_i, W_j, \gamma_0, \theta)] = 0$. The first question we address is whether the estimation of γ_0 impacts the asymptotic distribution of $\hat{\theta}^P$. We will answer this question in the affirmative, with the consequence that standard inference on the Gini coefficient is invalid, and regularization and model selection biases will also invalidate inference when high dimensional methods are used. In this paper we overcome these problems by using orthogonal (i.e. locally robust) quadratic moments for U-statistics and the corresponding debiased estimators. First, we introduce some general concepts needed before presenting the results for IOp.

2.2 General case

Consider the general case in which $g(w_i, w_j, \gamma, \theta)$ is some general vector with k known identifying moment functions. Without loss of generality, we can assume that g is a symmetric function in w_i and w_j .³ The terminology of *quadratic moment* comes from the fact that $\mathbb{E}[g(W_i, W_j, \gamma_0, \theta)]$ is a quadratic function of the true distribution F_0 , as

$$\mathbb{E}[g(W_i, W_j, \gamma_0, \theta)] = \int \int g(w_i, w_j, \gamma_0, \theta) F_0(dw_i) F_0(dw_j).$$

This feature is what differentiates our analysis from the standard debiased literature (see, e.g., Chernozhukov et al. (2022)). Since the computations for the Gini are complicated, we introduce a simpler running example to illustrate the main concepts and derivations. This example is of interest in its own.

EXAMPLE 1 (Variance of fitted values): Consider data $W_i = (Y_i, X_i)$, where Y_i is a real-valued outcome and X_i is a vector of covariates. The parameter of interest is the variance of the (population) fitted values $\theta_0 = \text{Var}(\gamma_0(X_i))$, which can be written as

$$\theta_0 = \mathbb{E} \left[\frac{(\gamma_0(X_i) - \gamma_0(X_j))^2}{2} \right].$$

This parameter is of interest for the analysis of variance in high dimensional settings and for IOp. The identifying moment function in this example is $g(w_i, w_j, \gamma, \theta) = (1/2)(\gamma(x_i) -$

³Otherwise, replace $g(w_i, w_j, \gamma_0, \theta)$ with its symmetrization $g^*(w_i, w_j, \gamma_0, \theta) = (1/2)[g(w_i, w_j, \gamma_0, \theta) + g(w_j, w_i, \gamma_0, \theta)]$.

$\gamma(x_j))^2 - \theta$. Solving an empirical analog of $\mathbb{E}[g(W_i, W_j, \gamma_0, \theta)] = 0$ we get the plug-in estimator

$$\hat{\theta}^P = \binom{n}{2}^{-1} \sum_{i < j} (1/2) (\hat{\gamma}(X_i) - \hat{\gamma}(X_j))^2.$$

This is the sample variance of fitted values. We show that $\hat{\theta}^P$ is highly biased when machine learning first steps are used. We develop inference methods for θ_0 and related parameters based on debiased estimators, which improve upon plug-in estimators. As an illustration, see Figure 1 where we simulate the plug-in and the debiased estimators for the variance of the fitted values example which we develop below. The Data Generating Process (DGP) is $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$ where the coefficients β_k , $k = 1, 2, 3$ are taken from a uniform distribution $U[0, 2]$, $\varepsilon \sim \mathcal{N}(0, 1/10)$ and $\text{Var}(X_i) = \Sigma$ with components $\Sigma_{ij} = 1(i = j) + (1/2) \times 1(|i - j| = 1)$. The fitted values are estimated with Random Forests. The histograms approximate the distributions of the centered estimators and the curves are normal p.d.f.s with the same variance as the estimators but with no bias (i.e. centered at zero). Even under this simple DGP and large sample sizes $n = 3000$ the plug-in estimator has a large bias compared to the debiased one. ■

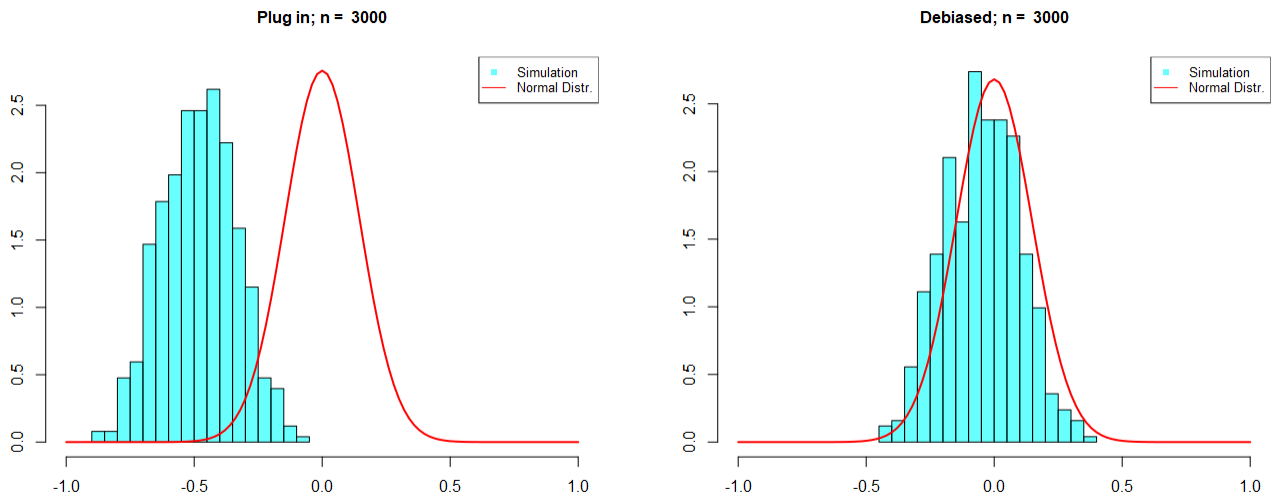


Figure 1: Comparison of plug-in and debiased estimator.

More broadly, plug-in estimators are generally biased by model selection and/or regularization in the first step for our U-statistic setting. The source of the bias problem is explained by the first term in the standard expansion of $U_n g(\cdot, \hat{\gamma}, \hat{\theta}^P)$ around θ_0 ,

$$0 = U_n g(\cdot, \hat{\gamma}, \theta_0) + \frac{\partial U_n g(\cdot, \hat{\gamma}, \bar{\theta}^P)}{\partial \theta'} (\hat{\theta}^P - \theta_0),$$

where $\bar{\theta}^P$ is an intermediate point between θ_0 and $\hat{\theta}^P$, and $U_n f = \binom{n}{2}^{-1} \sum_{i < j} f(X_i, X_j)$. The first term $U_n g(\cdot, \hat{\gamma}, \theta_0)$ can be further written as a standard U-statistic, $U_n g(\cdot, \gamma_0, \theta_0)$, and a first step estimation effect $U_n g(\cdot, \hat{\gamma}, \theta_0) - U_n g(\cdot, \gamma_0, \theta_0)$. The estimator $\hat{\theta}^P$ might inherit the regularization and model selection biases from $\hat{\gamma}$ through the term $U_n g(\cdot, \hat{\gamma}, \theta_0) - U_n g(\cdot, \gamma_0, \theta_0)$. Whether this is the case or not depends on the “derivative” of the mapping $\gamma \rightarrow \mathbb{E}[g(W_i, W_j, \gamma, \theta)]$ at γ_0 and whether this derivative is zero or not. For GMM moments this derivative is quantified by the First Step Influence Function (FSIF) introduced in [Chernozhukov et al. \(2022\)](#). However, the FSIF is not useful for providing a convenient quadratic representation of the impact of the first step, so we need a different representation than the standard one in [Chernozhukov et al. \(2022\)](#). We provide such quadratic representation in this paper and use it to construct the simple debiased U-estimator in our IOp application (cf. [1.1](#)).

2.2.1 The U-FSIF

As a first methodological contribution of this paper, we introduce the concept of *U-moment representation of the FSIF* (U-FSIF in short), ϕ , which takes into account the impact that the first step $\hat{\gamma}$ has on the estimation of the parameter of interest θ_0 . We claim the U-FSIF explains the observed bias in [Figure 1](#). To quantify such impact, let F be a cumulative distribution function (cdf) for W_i which is unrestricted except for regularity conditions such as the existence of $\gamma(F)$. For example, for $\gamma(F) = \mathbb{E}_F[Y|X]$ we require $\mathbb{E}_F[|Y|] < \infty$, where \mathbb{E}_F denotes expectation under F . Henceforth, $d/d\tau$ denotes the derivative from the right with respect to τ evaluated at $\tau = 0$ and set $F_\tau = F_0 + \tau(H - F_0)$ for some alternative distribution H and $\tau \in [0, 1]$. The alternative distribution H is chosen such that $\gamma_\tau = \gamma(F_\tau)$ exists for τ small enough. The U-FSIF ϕ is such that for all θ and all H

$$\int \int \phi(w_i, w_j, \gamma(F_\tau), \alpha(F_\tau), \theta) F_\tau(dw_i) F_\tau(dw_j) = 0 \text{ for all } \tau \in [0, \bar{\tau}), \bar{\tau} > 0, \quad (2.1)$$

and

$$\frac{d}{d\tau} \mathbb{E}[g(W_i, W_j, \gamma(F_\tau), \theta)] = \int \int \phi(w_i, w_j, \gamma_0, \alpha_0, \theta) K_H(dw_i, dw_j), \quad (2.2)$$

with $K_H(dw_i, dw_j) = F_0(dw_i)H(dw_j) + H(dw_i)F_0(dw_j)$. Here, α_0 is the α which satisfies [\(2.2\)](#). We may assume without loss of generality that ϕ is symmetric in w_i, w_j .⁴ The U-FSIF ϕ can be viewed as a quadratic representation of the influence function concept of [von Mises \(1947\)](#), [Hampel \(1974\)](#), [Newey \(1994\)](#) and [Ichimura and Newey \(2022\)](#). It can be found by

⁴As with g , setting $\phi^*(w_i, w_j, \gamma_0, \alpha_0, \theta) = (1/2)[\phi(w_i, w_j, \gamma_0, \alpha_0, \theta) + \phi(w_j, w_i, \gamma_0, \alpha_0, \theta)]$, we have $\int \int \phi(w_i, w_j, \gamma_0, \alpha_0, \theta) K_H(dw_i, dw_j) = \int \int \phi^*(w_i, w_j, \gamma_0, \alpha_0, \theta) K_H(dw_i, dw_j)$ and ϕ^* is symmetric.

solving the functional equation (2.2) for ϕ and it characterizes the local effect of the first step $\gamma(F)$ on the average moment functional $\mu(F) = \mathbb{E}[g(W_i, W_j, \gamma(F), \theta)]$ as F varies away from F_0 in any direction H . We give below in Section 3.1 new results characterizing ϕ for first steps satisfying orthogonality conditions.

EXAMPLE 1 (Variance of fitted values, cont.): By standard arguments

$$\frac{d}{d\tau} \mathbb{E} \left[\frac{(\gamma_\tau(X_i) - \gamma_\tau(X_j))^2}{2} - \theta \right] = \frac{d}{d\tau} \mathbb{E}[(\gamma_0(X_i) - \gamma_0(X_j))(\gamma_\tau(X_i) - \gamma_\tau(X_j))] \quad (2.3)$$

Define $\alpha_0(X_i, X_j) = \gamma_0(X_i) - \gamma_0(X_j)$, and use orthogonality and the chain rule to obtain

$$\frac{d}{d\tau} \mathbb{E}[\alpha_0(X_i, X_j)(Y_i - Y_j - \gamma_\tau(X_i) + \gamma_\tau(X_j))] = - \int \int \alpha_0(x_i, x_j)(y_i - y_j - \gamma_0(x_i) + \gamma_0(x_j)) K_H(dw_i, dw_j).$$

Thus, from the last two equations, we obtain a U-FSIF for this example

$$\phi(W_i, W_j, \gamma_0, \alpha_0, \theta) = \alpha_0(X_i, X_j) (Y_i - Y_j - \gamma_0(X_i) + \gamma_0(X_j)). \quad (2.4)$$

The function ϕ satisfies equations (2.1) and (2.2) for the g of this example, and explains the biased observed in Figure 1. ■

To motivate our definition of the U-FSIF, note that equation (2.1) is a zero mean condition and it implies by the chain rule

$$\begin{aligned} \frac{d}{d\tau} \mathbb{E}[\phi(W_i, W_j, \gamma(F_\tau), \alpha(F_\tau), \theta)] &= - \frac{d}{d\tau} \mathbb{E}_{F_\tau}[\phi(W_i, W_j, \gamma(F_0), \alpha(F_0), \theta)] \\ &= - \int \int \phi(w_i, w_j, \gamma_0, \alpha_0, \theta) K_H(dw_i, dw_j), \end{aligned} \quad (2.5)$$

for all θ and H . Equation (2.5) shows that the effect the first steps have on the U-FSIF “cancels out” with the effect they have on the original identifying moment in (2.2). Equation (2.5) also explains why K_H in the definition of a U-FSIF replaces the function H in the definition of the FSIF in Chernozhukov et al. (2022). The U-FSIF is not unique, however, this is not concerning since all U-FSIFs lead to the same first order asymptotics. The differences between the U-FSIF and the FSIF in Chernozhukov et al. (2022) and the (lack of) uniqueness of the U-FSIF are further explored in the Online Appendix.

Thus, letting $\psi(w_i, w_j, \gamma, \alpha, \theta) = g(w_i, w_j, \gamma, \theta) + \phi(w_i, w_j, \gamma, \alpha, \theta)$, we have an identification condition (by taking $\tau = 0$ in (2.1))

$$\mathbb{E}[\psi(W_i, W_j, \gamma_0, \alpha_0, \theta)] = 0 \text{ iff } \theta = \theta_0,$$

and local robustness with respect to first steps (by (2.2) and (2.5))

$$\frac{d}{d\tau} \mathbb{E}[\psi(W_i, W_j, \gamma(F_\tau), \alpha(F_\tau), \theta)] = 0.$$

This local robustness motivates estimators based on orthogonal moments. The defining property of the U-FSIF allows us to write θ as a solution to a locally robust *quadratic* moment function. This construction led to the expression of the new IOp estimator, which would not have been possible from the results in the debiased literature. In the general case, α_0 is an additional parameter on which the adjustment term ψ depends. In our running example and the leading IOp example this additional parameter is only a function of γ_0 , which substantially simplifies the construction of orthogonal quadratic moments and debiased estimators.

EXAMPLE 1 (Variance of fitted values, cont.): An orthogonal quadratic moment in this example is

$$\begin{aligned} \psi(W_i, W_j, \gamma_0, \alpha_0, \theta) &= (1/2)(\gamma_0(X_i) - \gamma_0(X_j))^2 - \theta + \alpha_0(X_i, X_j) (Y_i - Y_j - \gamma_0(X_i) + \gamma_0(X_j)) \\ &= (\gamma_0(X_i) - \gamma_0(X_j)) \left(Y_i - Y_j - \frac{\gamma_0(X_i) - \gamma_0(X_j)}{2} \right) - \theta, \end{aligned}$$

where recall $\alpha_0(X_i, X_j) = \gamma_0(X_i) - \gamma_0(X_j)$. ■

2.2.2 Debiased U-Estimators

For the estimation we have to take into account that estimation of nuisance parameters and moment conditions with the same observations can induce an “own-observation” bias. Also, machine learning first steps usually do not satisfy Donsker conditions. We use cross-fitting to overcome these issues (see [Bickel \(1982\)](#), [Schick \(1986\)](#), [Klaassen \(1987\)](#), [Chernozhukov et al. \(2018\)](#) and [Chernozhukov et al. \(2022\)](#)), but adapted to the U-statistics setting. First, we partition the observation indices $\mathcal{N} = \{1, \dots, n\}$ into K sets, $\mathcal{C} = \{C_1, \dots, C_K\}$. Then, \mathcal{C}^2 is a partition of all pairs of observations. Since we can focus on U-statistics with symmetric kernels, we use the partition of the set $\{(i, j) \in \mathcal{C}^2 : i < j\}$ into $L = K(K + 1)/2$ groups, $\mathcal{I} = \{I_1, \dots, I_L\}$. An illustration of this partition is given in the Online Appendix. Given a partition \mathcal{I} , let $\hat{\gamma}_l$ and $\hat{\alpha}_l$ be constructed using observations not present in the pairs in I_l . The debiased sample moment is

$$\hat{\psi}(\theta) = \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \left(g(W_i, W_j, \hat{\gamma}_l, \theta) + \phi(W_i, W_j, \hat{\gamma}_l, \hat{\alpha}_l, \theta) \right),$$

where $\hat{\alpha}_l$ is some estimator of α_0 which is discussed further in the Online Appendix. We call an estimator solving $\hat{\psi}(\hat{\theta}) = 0$ for a debiased/orthogonal moment a debiased U-estimator. We give sufficient conditions in the Appendix so that the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is $V = B^{-1}\Sigma B'^{-1}$, where

$$B = \frac{\partial \mathbb{E}[\psi(W_i, W_j, \gamma_0, \alpha_0, \theta)]}{\partial \theta}, \quad \Sigma = 4\text{Var}\left(\mathbb{E}[\psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)|W_i]\right).$$

The above asymptotic variance can be estimated as $\hat{V} = \hat{B}^{-1}\hat{\Sigma}\hat{B}'^{-1}$, where

$$\hat{B} = \binom{n}{2}^{-1} \sum_{i < j} \frac{\partial}{\partial \theta} \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \hat{\theta}),$$

$$\hat{\Sigma} = \frac{4}{n(n-1)^2} \sum_{i=1}^n \left[\sum_{j \neq i} \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \hat{\theta}) \right] \left[\sum_{j \neq i} \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \hat{\theta}) \right]'.$$

Note that we do not use cross-fitting to estimate the variance since we only care about its consistency, which simplifies the implementation and makes available standard U-statistics formulas for variances valid.

EXAMPLE 1 (Variance of fitted values, cont.): A debiased U-estimator for this example is

$$\hat{\theta} = \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} (\hat{\gamma}_l(X_i) - \hat{\gamma}_l(X_j)) \left(Y_i - Y_j - \frac{\hat{\gamma}_l(X_i) - \hat{\gamma}_l(X_j)}{2} \right).$$

This is the debiased U-estimator reported in Figure 1. To compute its standard errors, note first that $B = -1$. From the orthogonal moment function,

$$\mathbb{E}[\psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)|W_i] = \frac{1}{2} \left(\theta_0 + (\gamma_0(X_i) - \mu)^2 \right) + (\gamma_0(X_i) - \mu)\varepsilon_i,$$

where $\varepsilon_i = Y_i - \gamma_0(X_i)$ and $\mu = \mathbb{E}[Y_i]$. The variance term can be computed as

$$\Sigma = \text{Var}[(\gamma_0(X_i) - \mu)^2] + 4\mathbb{E}[(\gamma_0(X_i) - \mu)^2\varepsilon_i^2] + 4\mathbb{E}[(\gamma_0(X_i) - \mu)^3\varepsilon_i].$$

If $\gamma_0(X_i) = \mathbb{E}[Y_i|X_i]$ it simplifies to

$$\Sigma = \text{Var}[(\gamma_0(X_i) - \mu)^2] + 4\mathbb{E}[(\gamma_0(X_i) - \mu)^2\varepsilon_i^2].$$

Note that if γ_0 is non-constant, then $\Sigma > 0$ and the orthogonal moment function is a non-

degenerate kernel. The asymptotic variance can be estimated by

$$\hat{V} = \frac{4}{n(n-1)^2} \sum_{i=1}^n \left[\sum_{j \neq i} \left((\hat{\gamma}(X_i) - \hat{\gamma}(X_j)) \left(Y_i - Y_j - \frac{\hat{\gamma}(X_i) - \hat{\gamma}(X_j)}{2} \right) - \hat{\theta} \right) \right]^2.$$

Our theory ensures that the standard errors estimated with \hat{V} are robust to misspecification. This holds because the locally robust quadratic moments are computed from the limit $\gamma(F)$. Another popular parameter is the variance explained $R^2 = \theta_0 / \text{Var}[Y]$. The asymptotic theory and debiased estimator for the coefficient of determination in high dimensional settings follow directly from our results and a standard application of the delta method. These results are of independent interest. ■

2.3 Debiased IOp estimator

Now we introduce the IOp results. We focus here on the case in which $\hat{\gamma}$ is a nonparametric estimator of the conditional mean, i.e. $\gamma(F) = \mathbb{E}_F[Y|X]$. A first technical challenge we face for computing orthogonal moments in this example is the lack of differentiability of the absolute value. Despite this lack of differentiability, we are able to compute the U-FSIF from an application of our general results in the next section. Define $\Delta_0 = \gamma_0(X_i) - \gamma_0(X_j)$.

Assumption 1 (i) $\partial\gamma(F_\tau)/\partial\tau$ is bounded; either (ii) $\mathbb{P}[\Delta_0 = 0] = 0$ or (iii) $x_i \neq x_j \implies \gamma_0(x_i) - \gamma_0(x_j) \neq 0$.

Assumption 1(i) is a mild regularity assumption. Assumption 1(ii) is satisfied if $\gamma_0(X_i)$ is absolutely continuous, for example, if γ_0 is strictly monotone on a circumstance which is absolutely continuous given all the other circumstances. It is also implied by common margin assumptions in the classification literature which bound the rate at which the probability mass around zero decreases (e.g. [Mammen and Tsybakov \(1999\)](#) or [Tsybakov \(2004\)](#) among others). Assumption 1(iii) says that two observations with different circumstances must have different fitted values. Recall $\text{sgn}(x) = 1(x > 0) - 1(x < 0)$.

Proposition 1 Under Assumption 1, the U-FSIF in the Gini of fitted values is given by

$$\phi(w_i, w_j, \gamma_0, \theta) = \theta(y_i + y_j - \gamma_0(x_i) - \gamma_0(x_j)) - \alpha_0(X_i, X_j)(y_i - y_j - \gamma_0(x_i) + \gamma_0(x_j)),$$

where $\alpha_0(x_i, x_j) = \text{sgn}(\gamma_0(x_i) - \gamma_0(x_j))$.

Let $\hat{\alpha}_i(X_i, X_j) = \text{sgn}(\hat{\gamma}_i(x_i) - \hat{\gamma}_i(x_j))$. Adding the U-FISF to the identifying moment, sim-

plifying and solving the debiased orthogonal sample moment we get the debiased estimator

$$\begin{aligned}\hat{\theta} &= \frac{\sum_{l=1}^L \sum_{(i,j) \in I_l} |\hat{\gamma}_l(X_i) - \hat{\gamma}_l(X_j)| + \hat{\alpha}_l(X_i, X_j)(Y_i - Y_j - \hat{\gamma}_l(X_i) + \hat{\gamma}_l(X_j))}{\sum_{i < j} (Y_i + Y_j)} \\ &= \frac{\sum_{l=1}^L \sum_{(i,j) \in I_l} \hat{\alpha}_l(X_i, X_j)(Y_i - Y_j)}{\sum_{i < j} (Y_i + Y_j)}.\end{aligned}\tag{2.6}$$

The nonparametric debiased IOp estimator resembles the standard Gini coefficient for income, but rather than weighting by the sign of the differences in income, the debiased IOp estimator weights by the sign of the difference in fitted values. This means that whenever two individuals have the same fitted values their difference in incomes cannot be attributed to inequality of opportunity. To give the expression for the asymptotic variance, define $d(u, t) = \mathbb{E}[\text{sgn}(u - \gamma_0(X_j))(t - Y_j)]$. Since the asymptotic variance of a U-statistic with kernel $g(X_i, X_j)$ is given by $4\text{Var}(\mathbb{E}[g(X_i, X_j)|X_i])$ and

$$\mathbb{E}[\psi(W_i, W_j, \gamma_0, \alpha_0, \theta)|W_i] = \theta(Y_i + \mu) - \mathbb{E}[d(\gamma_0(X_i), Y_i)],$$

we have that the asymptotic variance of the debiased estimator is given by

$$\Sigma = 4\text{Var}\left(\theta_0 Y_i - d(\gamma_0(X_i), Y_i)\right).$$

If $\Sigma > 0$ then the asymptotic variance of $\hat{\theta}$ can be estimated by

$$\hat{V} = \frac{\frac{1}{n(n-1)^2} \sum_{i=1}^n \left(\sum_{j \neq i} \hat{\theta}(Y_i + Y_j) - \hat{\alpha}_l(X_i, X_j)(Y_i - Y_j) \right)^2}{\left(n^{-1} \sum_{i=1}^n Y_i \right)^2}.$$

In the next section we provide conditions that guarantee valid asymptotic inference for the IOp parameter θ_0 based on $\hat{\theta}$ and the variance estimate \hat{V} .

2.4 Asymptotic properties of the debiased IOp

We give low-level conditions for the IOp example based on the general asymptotic properties given in the Online Appendix. Define the difference of fitted values $\Delta_\gamma(X_i, X_j) = \gamma(X_i) - \gamma(X_j)$ and recall $\Delta_0 \equiv \Delta_{\gamma_0}(X_i, X_j)$. To obtain asymptotic normality assume the following

Assumption 2 (i) $\mathbb{E}[(Y_i - \gamma_0(X_i))^2] \leq C$;

(ii) $\int |\hat{\gamma}_l(x) - \gamma_0(x)|^2 F_0(w) \rightarrow_p 0$; (iii) $\int \int |\hat{\alpha}_l(x_i, x_j) - \alpha_0(x_i, x_j)|^2 F_0(dw_i) F_0(dw_j) \rightarrow_p 0$;

$$(iv) \sqrt{n} \int \int (\hat{\alpha}_l(x_i, x_j) - \alpha_0(x_i, x_j))(\hat{\gamma}_l(x_i) - \gamma_0(x_i) - \hat{\gamma}_l(x_j) + \gamma_0(x_j)) F_0(dw_i) F_0(dw_j) \rightarrow_p 0.$$

Assumption 2 (i) imposes a mild boundedness condition. Assumptions 2 (ii) and (iii) require $\hat{\gamma}_l$ and $\hat{\alpha}_l$ to be mean square consistent, which are also mild conditions and allow for generic estimators (for conditions on specific estimators see the next paragraphs). In the next proposition we give conditions under which (iii) and (iv) hold. Let $a_n = \|\hat{\gamma}_l - \gamma_0\|_\infty = \sup_{x \in \mathcal{X}} |\hat{\gamma}_l(x) - \gamma_0(x)|$.

Proposition 2 *Suppose that any of the following holds*

(i) $a_n = o_p(n^{-1/4})$, the c.d.f. of Δ_0 is absolutely continuous and its density is bounded around zero,

(ii) there exists $\eta > 0$ such that $\mathbb{P}(|\Delta_0| \leq \eta \mid X_i \neq X_j) = 0$ and $\|\hat{\gamma}_l - \gamma_0\| = o_p(n^{-1/4})$.

Then, Assumptions 2 (iii)-(iv) follow.

Condition (i) in Proposition 2 strengthens the convergence of $\hat{\gamma}_l$ to uniform convergence, requires absolute continuity of the c.d.f. of Δ_0 and asks for a bounded density. These are standard conditions in the nonparametric literature. Condition (ii) weakens the uniform convergence to mean square convergence, at the cost of ruling zero as an element of the support of Δ_0 . This assumption is realistic whenever circumstances are discrete and Assumption 1 (iii) holds. For instance if $X_i \in \{0, 1\}$ and $\gamma_0(0) \neq \gamma_0(1)$, then condition (ii) follows. Condition (ii) is a plausible assumption for our empirical application where all circumstances are discrete.

There is a large literature checking L_2 -convergence rates for different machine learners under low level sparsity or smoothness conditions on the nuisance parameters. The traditional non-parametric literature gives rates for kernel regression and sieves/series (e.g. [Chen \(2007\)](#)). For L_1 -penalty estimators such as Lasso see, e.g., [Belloni and Chernozhukov \(2011\)](#) and [Belloni and Chernozhukov \(2013\)](#). Also for low level conditions for shrinkage and kernel estimators see Appendix B in [Sasaki and Ura \(2021\)](#). Rates for L_2 -boosting in low dimensions are found in [Zhang and Yu \(2005\)](#), and more recently [Kueck et al. \(2023\)](#) find rates for L_2 -boosting with high dimensional data. For results on versions of random forests see [Wager and Walther \(2015\)](#) and [Athey et al. \(2019\)](#). Finally, for single-layer, sigmoid-based neural networks see [Chen and White \(1999\)](#) and for a modern setting of deep neural networks with rectified linear (ReLU) activation function see [Farrell et al. \(2021\)](#). For instance, Theorem 1 in [Farrell et al. \(2021\)](#) shows that for $X \in [-1, 1]^d$ continuously distributed and bounded Y and $\gamma_0(X)$, we have that with high probability for a deep ReLU network estimator $\hat{\gamma}$ for large n

$$\|\hat{\gamma} - \gamma\|^2 \leq C \left(n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log \log n}{n} \right),$$

where β is a smoothness parameter and C is some positive constant. Also, under some conditions, [Kueck et al. \(2023\)](#) give a L_2 -rate of $s \log(d \vee n) \log n/n$ for L_2 -boosting, which almost attains the usual Lasso rate $s \log(d \vee n)/n$, where d is the dimension of X and s is a sparsity parameter. For uniform convergence rates, see, e.g., [Hansen \(2008\)](#) for kernel regression and [Lounici \(2008\)](#) or [van der Laan and Bibaut \(2017\)](#) for Lasso.

Now we move to the conditions for consistent estimation of the variance. Since cross-fitting is not needed to show the consistency of the variance, we repeat some of the assumptions in [Assumption 2](#) for nuisance estimators which use all observations.

Assumption 3 (i) $\mathbb{E}[(\hat{\gamma}(X_i) - \gamma_0(X_i))^2] \rightarrow 0$; (ii) $\mathbb{E}[(\hat{\alpha}(X_i, X_j) - \alpha_0(X_i, X_j))^2] \rightarrow 0$.

Proposition 3 Under [Assumptions 2](#) and [3](#) and $\hat{\theta} \rightarrow_p \theta_0$, we have that $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d \mathcal{N}(0, V)$ where $V = \text{Var}\left(\mathbb{E}[\psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)|W_i]\right)/(4\mathbb{E}[Y_i]^2)$ and $\hat{V} \rightarrow_p V$.

Consistency of $\hat{\theta}$ follows readily from the mild conditions given in the Online Appendix.

3 General theory for semiparametric U-statistics

This section covers the theory of constructing locally robust estimators for semiparametric U-statistics and their asymptotic properties in full generality. This section is of interest to the reader interested in the technical details of the method and the practitioner who wishes to know how to construct a debiased estimator for a semiparametric U-statistic of his/her choice. The reader whose main interest is the IOp estimator can safely skip these results.

3.1 The first step and the construction of the U-FSIF

The expression of the U-FSIF ϕ depends on the original identifying moment g and the first step limit $\gamma(F)$. We consider first step functions in a linear set Γ which satisfy a linear orthogonality condition, as this setting is quite general and fits well the applications we are interested in. Suppose $W_i = (Y_i, X_i) \in \mathcal{Y} \times \mathcal{X}$ and let $\gamma(F) \equiv \gamma_F(X_i)$, $\gamma(F) \in \Gamma$, be such that

$$\mathbb{E}_F[\nu(X_i)(Y_i - \gamma(F))] = 0 \text{ for all } \nu \in \Gamma. \quad (3.1)$$

This setting covers high dimensional and nonparametric regression, additive regression and single index models, among others (see [Ichimura and Newey \(2022\)](#)). We assume that $\Gamma \subseteq L_2(X_i)$, where $L_2(X_i)$ is the set of squared integrable functions of X_i (similarly, we define $L_2(X_i, X_j)$). Sometimes we drop the reference to random variables and use simply L_2 (the

meaning will be clear from the context). For simplicity of exposition, we take $\gamma(F)$ to be real-valued, though the extension of our results to multiple first steps follows straightforwardly from the chain rule.

An example that has been popular in machine learning, but less explored in econometrics, is kernel machine regression, where the first step solves

$$\hat{\gamma} = \arg \min_{\gamma \in H_K} \frac{1}{n} \sum_{i=1}^n [Y_i - \gamma(X_i)]^2 + \lambda \|\gamma\|_{H_K}.$$

Here, $\Gamma = H_K$ is a reproducing kernel Hilbert space with kernel K (see, e.g., [Scholkopf and Smola \(2018\)](#)), $\|\cdot\|_{H_K}$ its norm, and λ is a penalization parameter converging to zero with the sample size. By the Representer Theorem, see Theorem 4.2 in [Scholkopf and Smola \(2018\)](#), there is a closed form expression for $\hat{\gamma}$, and its limit satisfies (3.1) with $\Gamma = H_K$. Therefore, kernel machine regression is also included in our setting as a special case.

The starting point of our analysis is the linearization step of the original identifying moment function g : assume there exists a function $\delta \in L_2$ and constants c_1 and c_2 such that

$$\frac{d}{d\tau} \mathbb{E}[g(W_i, W_j, \gamma(F_\tau), \theta)] = \frac{d}{d\tau} \mathbb{E}[\delta(X_i, X_j, \gamma_0)(c_1 \gamma_\tau(X_i) + c_2 \gamma_\tau(X_j))], \quad (3.2)$$

where $\gamma_\tau(x) = \gamma(F_\tau)(x)$ denotes the orthogonal projection pertaining to F_τ . For simplicity of notation, we drop the possible dependence of δ , c_1 and c_2 on θ and also the possible dependence of c_1 and c_2 on γ_0 . Henceforth, we also use the short notation $\delta_{ij}(\gamma) \equiv \delta(X_i, X_j, \gamma)$.

We illustrate this preliminary linearization step in the variance of fitted values example.

EXAMPLE 1 (Variance of fitted values, cont.): As we have seen in (2.3), (3.2) holds for this example with $\delta_{ij}(\gamma_0) = \gamma_0(X_i) - \gamma_0(X_j)$ and $(c_1, c_2) = (1, -1)$. ■

After the linearization, the second step is a projection step. This step exploits that $c_1 \gamma_0(X_i) + c_2 \gamma_0(X_j)$ can be viewed as a projection of $c_1 Y_i + c_2 Y_j$ onto a closed set \mathcal{S} , $\mathcal{S} \subseteq L_2(X_i, X_j)$, satisfying some regularity conditions. We assume that \mathcal{S} is a closed linear set of functions of (X_i, X_j) containing $\Gamma + \Gamma := \{v(X_i) + w(X_j) : v, w \in \Gamma\}$. We also assume that Γ is a closed linear set of functions that contains constant functions. Let $\Pi_V(\cdot)$ be the orthogonal projection operator onto V , for a closed linear set V . Define $\alpha_0(X_i, X_j) \equiv \Pi_{\mathcal{S}}(\delta(X_i, X_j, \gamma_0))$, where δ is as in (3.2).

Lemma 1 *Suppose (3.1) and (3.2) hold. If α_0 is such that $\alpha_0(\cdot, x) \in \Gamma$ and $\alpha_0(x, \cdot) \in \Gamma$ for all x , then (2.2) holds with*

$$\phi(W_i, W_j, \gamma_0, \alpha_0, \theta) = \alpha_0(X_i, X_j) (c_1 Y_i + c_2 Y_j - c_1 \gamma_0(X_i) - c_2 \gamma_0(X_j)). \quad (3.3)$$

Note that (2.1) is also satisfied.

This lemma gives some flexibility in the choice of \mathcal{S} , as long as it satisfies the stated conditions. Thus, we could apply Lemma 1 to different choices of \mathcal{S} to obtain different weights $\alpha_0(X_i, X_j) \equiv \Pi_{\mathcal{S}}(\delta(X_i, X_j, \gamma_0))$ and the corresponding U-FSIFs. However, in applications where $\delta_{ij}(\gamma_0)$ is linear in $\gamma_0(X_i)$ and $\gamma_0(X_j)$, as in the variance of fitted values example, we have that $\alpha_0 = \delta(\cdot, \gamma_0)$, independently on the choice of $\mathcal{S} \supseteq \Gamma + \Gamma$, and hence, in these cases the expression of α_0 is unique.

EXAMPLE 1 (Variance of fitted values, cont.): Since $\delta_{ij}(\gamma) = \gamma(X_i) - \gamma(X_j)$, $\gamma \in \Gamma$, and $\mathcal{S} \supseteq \Gamma + \Gamma$, it follows that $\alpha_0(X_i, X_j) \equiv \Pi_{\mathcal{S}}(\delta(X_i, X_j, \gamma_0)) = \delta(X_i, X_j, \gamma_0)$, independently of \mathcal{S} , and hence

$$\alpha_0(X_i, X_j) = \gamma_0(X_i) - \gamma_0(X_j).$$

This is true regardless of the set Γ . Thus, for this example, the U-FSIF from Lemma 1 is given by

$$\phi(W_i, W_j, \gamma_0, \alpha_0, \theta) = \alpha_0(X_i, X_j) (Y_i - Y_j - \gamma_0(X_i) + \gamma_0(X_j)). \quad (3.4)$$

■

More generally, the following result provides two instances where the conditions of Lemma 1 are satisfied for the given Γ and \mathcal{S} and the different expressions of α_0 provided. To introduce the result, we define $\alpha_{01}(X_i) = \Pi_{\Gamma} \mathbb{E}[\delta_{ij}(\gamma_0) | X_i]$ and $\alpha_{02}(X_j) = \Pi_{\Gamma} \mathbb{E}[\delta_{ij}(\gamma_0) | X_j]$.

Lemma 2 Suppose (3.1) and (3.2) hold. Then:

(i) For $\Gamma = L_2(X_i)$, if $\mathcal{S} = L_2(X_i, X_j)$, then $\alpha_0 = \delta(\cdot, \gamma_0)$.

(ii) For any Γ , if $\mathcal{S} = \Gamma + \Gamma$, then

$$\alpha_0(X_i, X_j) = \alpha_{01}(X_i) + \alpha_{02}(X_j) - \mathbb{E}[\delta_{ij}(\gamma_0)]. \quad (3.5)$$

We refer henceforth to the situation of Lemma 2(i) as the joint nonparametric case. The Γ and \mathcal{S} in (i) lead to the simple expression for the debiased estimator in our application to IOp given in (1.1). Lemma 2(ii) is most useful in settings where considering joint nonparametric estimation is not feasible (e.g. in high dimensional settings), and thus the researcher chooses as Γ a strict subset of $L_2(X_i)$.

We extend the scope of applications of Lemmas 1 and 2 to cases where for some $M \geq 1$,

$$\frac{d}{d\tau} \mathbb{E}[g(W_i, W_j, \gamma(F_\tau), \theta)] = \sum_{m=1}^M \frac{d}{d\tau} \mathbb{E}[\delta_m(X_i, X_j, \gamma_0, \theta) (c_{1m} \gamma_\tau(X_i) + c_{2m} \gamma_\tau(X_j))], \quad (3.6)$$

for functions $\delta_m \in L_2$ and constants c_{1m} and c_{2m} . Applying Lemma 1 to each of the summands in (3.6) we obtain the U-FSIF

$$\phi(W_i, W_j, \gamma_0, \alpha_0, \theta) = \sum_{m=1}^M \alpha_{0m}(X_i, X_j) (c_{1m}Y_i + c_{2m}Y_j - c_{1m}\gamma_0(X_i) - c_{2m}\gamma_0(X_j)), \quad (3.7)$$

where $\alpha_{0m}(X_i, X_j) \equiv \Pi_S(\delta_m(X_i, X_j, \gamma_0))$. In particular, we can exploit the additivity of expectations in (3.2) and write

$$\frac{d}{d\tau} \mathbb{E}[g(W_i, W_j, \gamma(F_\tau), \theta)] = c_1 \frac{d}{d\tau} \mathbb{E}[\delta(X_i, X_j, \gamma_0)\gamma_\tau(X_i)] + c_2 \frac{d}{d\tau} \mathbb{E}[\delta(X_i, X_j, \gamma_0)\gamma_\tau(X_j)]$$

to obtain the alternative U-FSIF

$$\phi(W_i, W_j, \gamma_0, \alpha_0, \theta) = c_1 \alpha_{01}(X_i)(Y_i - \gamma_0(X_i)) + c_2 \alpha_{02}(X_j)(Y_j - \gamma_0(X_j)), \quad (3.8)$$

which corresponds to (3.7) with $M = 2$, $c_{11} = c_1$, $c_{21} = 0$, $c_{12} = 0$, $c_{22} = c_2$, $\alpha_{01}(X_i, X_j) \equiv \alpha_{01}(X_i)$ and $\alpha_{02}(X_i, X_j) \equiv \alpha_{02}(X_j)$.

EXAMPLE 1 (Variance of fitted values, cont.): Applying the expression in (3.8) to this example we obtain a U-FSIF given by

$$\phi(W_i, W_j, \gamma_0, \alpha_0, \theta) = \alpha_{01}(X_i)(Y_i - \gamma_0(X_i)) - \alpha_{02}(X_j)(Y_j - \gamma_0(X_j)), \quad (3.9)$$

where $\alpha_{01}(x) = \gamma_0(x) - \mathbb{E}[Y_i]$ and $\alpha_{02}(x) = \mathbb{E}[Y_i] - \gamma_0(x)$. ■

The representation in (3.8) is also useful for giving conditions under which there is no first step estimation effect in standard errors, see Remark 1. The definition of the FSIF is given in Chernozhukov et al. (2022).

Lemma 3 Suppose (3.1) and (3.2) hold. Then, the FSIF is given by $2\phi_1(W_i, \gamma, \alpha)$ where

$$\phi_1(W_i, \gamma, \alpha) = \frac{1}{2} [c_1 \alpha_{01}(X_i) + c_2 \alpha_{02}(X_i)] \times (Y_i - \gamma_0(X_i)).$$

Thus, there is no estimation effect from first steps if $c_1 \alpha_{01}(x) + c_2 \alpha_{02}(x) \equiv 0$.

EXAMPLE 1 (Variance of fitted values, cont.): The corresponding FSIF from Lemma 3 is $2\phi_1$ where

$$\begin{aligned} \phi_1(W_i, \gamma, \alpha) &= \frac{1}{2} [\gamma_0(X_i) - \mathbb{E}[Y_i] - (\mathbb{E}[Y_i] - \gamma_0(X_i))] \times (Y_i - \gamma_0(X_i)) \\ &= (\gamma_0(X_i) - \mathbb{E}[Y_i])(Y_i - \gamma_0(X_i)). \end{aligned}$$

As discussed in the Online Appendix, an important difference with respect to the debiased literature is that the adjustment term in our case is not unique. However, any U-FSIF leads to the same first order asymptotics. In fact, the U-FSIF in (3.9) coincides with the symmetrized version of the FSIF for this problem, and the difference between (3.9) and the U-FSIF given earlier in (3.4) is the degenerate kernel

$$\xi(w_i, w_j) = (\mathbb{E}[Y_j] - \gamma_0(x_j))(y_i - \gamma_0(x_i)) - (\mathbb{E}[Y_i] - \gamma_0(x_i))(y_j - \gamma_0(x_j)).$$

■

We finish this section by showing local robustness with respect to γ and global robustness with respect to α from the result in Lemma 1. Because $\alpha_0 = \Pi_{\mathcal{S}}\delta$ and $\Gamma + \Gamma \subseteq \mathcal{S}$, the local first step effect on the identifying moment from (3.2) equals (by iterated projections) to

$$\frac{d}{d\tau}\mathbb{E}[\delta(X_i, X_j, \gamma_0)(c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j))] = \frac{d}{d\tau}\mathbb{E}[\alpha_0(X_i, X_j)(c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j))],$$

which cancels out with the local effect of γ on the average of ϕ , i.e.

$$\frac{d}{d\tau}\mathbb{E}[\alpha_0(X_i, X_j)(c_1Y_i + c_2Y_j - c_1\gamma_\tau(X_i) - c_2\gamma_\tau(X_j))] = -\frac{d}{d\tau}\mathbb{E}[\alpha_0(X_i, X_j)(c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j))].$$

Global robustness about α holds because by orthogonality, for all $\alpha \in \mathcal{S}$,

$$\mathbb{E}[\alpha(X_i, X_j)(c_1Y_i + c_2Y_j - c_1\gamma_0(X_i) - c_2\gamma_0(X_j))] = 0.$$

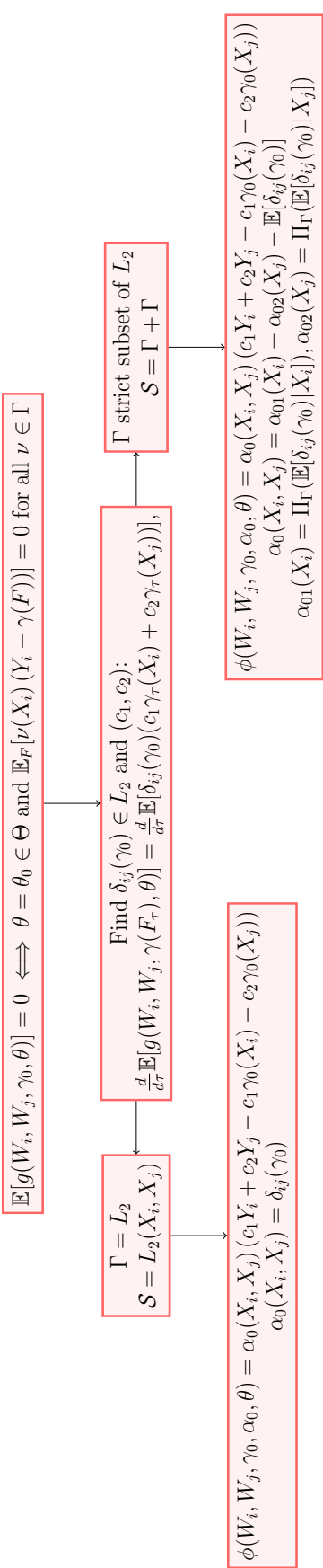
This global robustness property on α implies that only weak conditions about the convergence of machine learning estimators for α_0 will be required in our asymptotic results.

EXAMPLE 1 (Variance of fitted values, cont.): Global robustness follows since for $\alpha \in \Gamma + \Gamma$ and by orthogonality

$$\mathbb{E}[\alpha(X_i, X_j)(Y_i - Y_j - \gamma_0(X_i) + \gamma_0(X_j))] = 0.$$

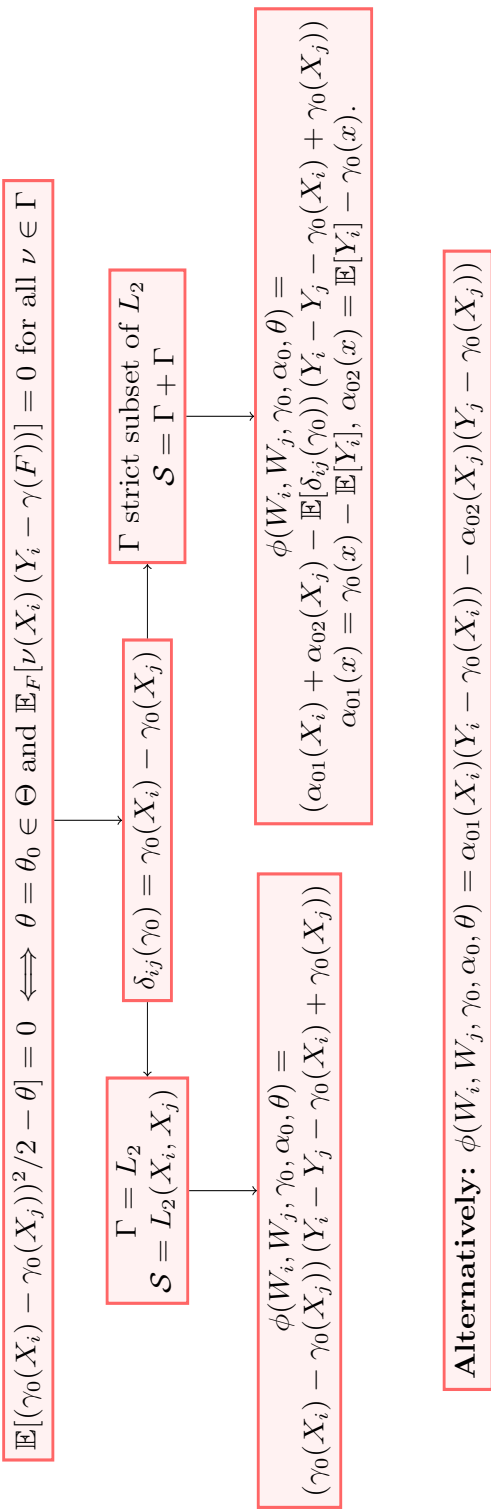
■

For a practitioner with a given semiparametric quadratic moment in mind we ease the process of finding the U-FSIF by summing up the whole process for the general case and for our running example in Figures 2 and 3.



Alternatively: $\phi(W_i, W_j, \gamma_0, \alpha_0, \theta) = c_1 \alpha_{01}(X_i)(Y_i - \gamma_0(X_i)) + c_2 \alpha_{02}(X_j)(Y_j - \gamma_0(X_j))$

Figure 2: How to find U-FSIFs for general quadratic moment condition.



Alternatively: $\phi(W_i, W_j, \gamma_0, \alpha_0, \theta) = \alpha_{01}(X_i)(Y_i - \gamma_0(X_i)) - \alpha_{02}(X_j)(Y_j - \gamma_0(X_j))$

Figure 3: How to find U-FSIFs for the variance of the fitted values.

4 Simulations

We evaluate our debiased estimator for the Gini of the fitted values with Monte Carlo simulations. We consider three independent categorical covariates with eight equally likely levels, $X_{k,i} \in \{0, 1, \dots, 7\}$ for $k = 1, 2, 3$. Define dummy variables $D_{kj,i} = 1(X_{k,i} = j)$ for $j = 0, 1, \dots, 7$. We use the following DGP:

$$\ln(Y_i) = \beta_0 + \sum_{k=1}^3 \sum_{j=1}^7 \beta_{k,j} D_{kj,i} + \sum_{k < k'} \sum_{r=1}^7 \sum_{s=1}^7 \delta_{kk',rs} D_{kr,i} D_{k's,i} + \sum_{r=1}^7 \sum_{s=1}^7 \sum_{t=1}^7 \gamma_{rst} D_{1r,i} D_{2s,i} D_{3t,i} + \varepsilon_i.$$

The above constitutes a saturated model with a reference group represented by the constant, $3 \times 7 = 21$ coefficients of main effects, $\binom{3}{2} \times 7^2 = 147$ coefficients for pairwise interactions among the levels and $7^3 = 343$ coefficients for threewise interactions among the levels. Hence, in total we have 512 parameters which means that estimating this model for low sample sizes is a high dimensional problem. The noise variable ε_i is distributed as $\mathcal{N}(0, \sigma^2)$. We set $\beta_0 = 5$ and for the vector $\beta = (\beta_{11}, \dots, \beta_{1,7}, \beta_{2,1}, \dots, \beta_{3,7})'$, we set $\beta_m = 0.2(-1)^{m+1}$ for $m = 1, \dots, 21$. For the vector of interaction coefficients $\xi = (\delta', \gamma)'$ we set $\xi_m = (2m^2)^{-1}$ for $m = 1, \dots, 490$.

In the simulations we set the variance of the unobservable term σ^2 to 0.1. This implies a Signal to Noise ratio of $StN = 10.5$. For the results for $\sigma \in \{0.2, 0.3\}$ (and correspondingly $StN \in \{2.6, 1.2\}$) see the Online Appendix. To estimate the fitted values we run Lasso and Ridge regressions of $\ln(Y_i)$ using 10-fold Cross-Validation for the regularization parameters. For the IOp we apply the exponential function to the fitted values of the Lasso and Ridge regressions and then use the plug-in and debiased estimators since we are interested in the IOp of Y_i and not of $\ln(Y_i)$.⁵ For the debiased estimator we partition the observation indices into 5 sets for cross-fitting, i.e. $K = 5$. We also estimate the first step with Random Forests (RF), Conditional Inference Forests (CIF), Extreme Gradient Boosting (XGB) and Catboost (CB). The results are in Table 1.

The debiased estimators are unstable for $n = 100$. This is expected since the first steps are using very few observations due to the cross-fitting procedure. This issue disappears fast as n increases. In the case of Lasso we see that for sample sizes larger than 100 the bias of the debiased estimators is considerably smaller in absolute size than the plug-in biases. Also, the coverage rate of the debiased estimator is much closer to the nominal 95% compared to the coverage of the plug-in estimator. In the case of Ridge there is not much difference between the plug-in estimator and the debiased estimator, possibly due to the weak sparsity of the model, and the coverage of both confidence intervals slowly increases to the correct

⁵Under a log-normal model $\ln(Y_i) = \beta'X_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, we have that $\mathbb{E}[Y_i|X_i] = e^{\beta'X_i + \sigma^2/2}$, hence the fitted values $\gamma(x) = e^{\beta'x}$ solve $\mathbb{E}[\nu(X_i)(Y_i/e^{\sigma^2/2} - \gamma(X_i))] = 0$ for all $\nu(\cdot) \in L_2$. Hence, we can use the estimator of the joint nonparametric case and the constant $e^{\sigma^2/2}$ cancels.

Gini of the FVs								
	plug-in Lasso		Debiased Lasso		plug-in Ridge		Debiased Ridge	
	Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage
n = 100	-0.002	0.863	-0.026	0.567	-0.086	0.000	-0.022	0.629
n = 500	-0.005	0.746	-0.002	0.909	-0.081	0.000	-0.005	0.798
n = 1000	-0.005	0.710	-0.001	0.950	-0.002	0.877	-0.003	0.879
n = 3000	-0.002	0.752	0.000	0.940	0.000	0.927	-0.001	0.931
	plug-in RF		Debiased RF		plug-in CIF		Debiased CIF	
	Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage
n = 100	-0.032	0.157	-0.015	0.774	-0.072	0.002	-0.017	0.730
n = 500	-0.022	0.008	-0.002	0.917	-0.034	0.000	-0.003	0.923
n = 1000	-0.022	0.000	-0.001	0.937	-0.027	0.000	-0.002	0.923
n = 3000	-0.023	0.000	-0.001	0.933	-0.023	0.000	-0.001	0.942
	plug-in XGBoost		Debiased XGBoost		plug-in Catboost		Debiased Catboost	
	Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage
n = 100	0.010	0.851	-0.039	0.310	-0.012	0.619	-0.032	0.429
n = 500	0.006	0.774	-0.005	0.863	0.002	0.815	-0.003	0.907
n = 1000	0.004	0.772	-0.002	0.931	0.002	0.837	-0.001	0.929
n = 3000	0.003	0.702	0.000	0.956	0.001	0.829	0.000	0.956

Table 1: Simulation based on 500 Monte Carlo iterations, true value for the Gini of the FVs is 0.18.

nominal size as the sample size increases.

The performance of RF and CIF is similar. The biases of the debiased estimators are much lower in absolute value than that of the plug-in estimators. Also, the coverage rates of the plug-in estimators are far from the nominal values while the coverage of the debiased estimators are close to the nominal value. Finally, the boosting methods XGB and CB show little bias for the debiased and plug-in estimators. However, only the debiased estimators achieve coverage rates close to the nominal values.

Hence, in all cases the debiased estimators achieve little bias and are the only estimators which allow for correct inference.

5 Inequality of Opportunity in Europe

We measure IOp in 29 European countries using our debiased IOp estimator and the 2019 wave of EUSILC. Our measure of income is equalized household income and the level of

observation is the individual.⁶ We restrict the sample to those aged between 25 and 59 years old to focus on the working age population.

The circumstances included in the intergenerational module include questions on characteristics of the parents and questions related to the individual's life/household when he/she was around 14 years old. We use the following circumstances: sex, country of birth, whether he/she was living with the mother/father, the number of adults/working adults/kids in the household, population of the municipality, tenancy of the house, country of birth of the parents, nationality of the parents, education of the parents, occupational status of the parents, father's managerial position, father's occupation, basic school needs (whether he/she had access to books, materials, etc.), financial situation, food needs (whether he/she could eat meat/chicken/fish/vegetarian equivalent once a week and holidays outside of home once a year). Remember that it refers to when the individual was around 14 years old, so, for instance, financial situation refers to the financial situation of the household where the individual resided when he/she was around 14 years old.

All circumstances are discrete and there are many different combinations of the categories. This makes the problem high dimensional and hard to deal with without machine learning procedures. In this application we use Lasso, Ridge, RF, CIF, XGB and CB. For Lasso and Ridge we use dummy encoding and employ a dictionary including up to 8-wise interactions (1,658 regressors). A nice feature of the cross-fitting procedure is that one can retrieve a cross-validated estimate of the Root Mean Square Error (RMSE) in the first stage. We will show our results for each country only for the machine learner which attains the lowest RMSE when predicting income in the first stage; we call this machine learner the best performing machine learner.

For cross-fitting we split $\{1, \dots, n\}$ into C_1, \dots, C_K with $K = 5$ which leads to $L = K(K + 1)/2$ blocks I_1, \dots, I_L (see Online Appendix for details). The debiased and plug-in relative IOp estimates (IOp as a fraction of the Gini of income) based on the best performing machine learner with 95 % confidence intervals for the debiased estimates can be seen in Figure 4.

We see a lot of heterogeneity in IOp across different European countries. Relative IOp, i.e. Gini of the fitted values over the Gini of income, takes values from 5% to almost 60%. As usual, Nordic countries such as Denmark, Finland or Norway are among the countries with the lowest relative IOp with 5%, 13% and 24% of inequality explained by circumstances respectively. In Denmark, predicting income from circumstances seems to be a formidable task, suggesting that income is close to mean independent from the observed circumstances. From the Nordic countries, the one with the highest relative IOp is Sweden with 31% of relative IOp. Netherlands and Germany are also in the lower range with relative IOp of 18%

⁶This variable is the total household disposable income in 2018 (variable HY020) divided by the equivalence scale given in the database (HX240) which is the modified OECD equivalence scale.

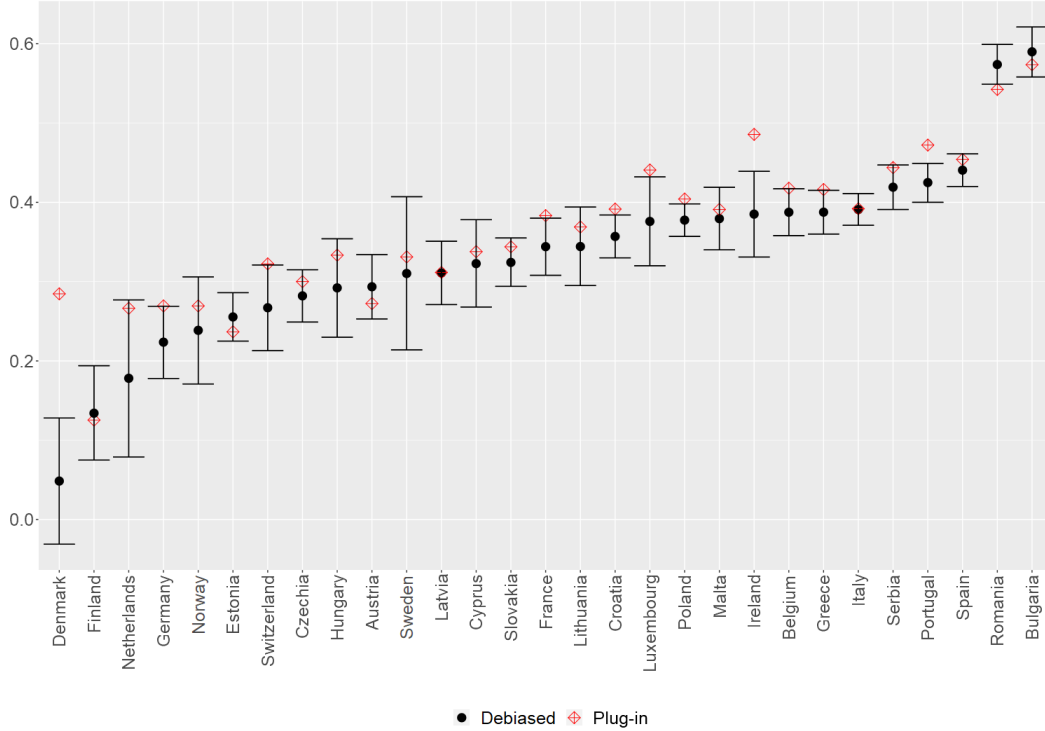


Figure 4: Gini IOP in Europe

and 22% respectively. Eastern countries are more heterogeneous, in the lower range we have countries such as Czechia or Hungary with relative IOP of 28% and 29% respectively. In the higher range, we have Romania and Bulgaria with relative IOP being almost 60%. Southern countries generally have high levels of IOP; Greece, Italy and Spain have 39%, 39% and 44% of total inequality explained by circumstances.

Comparing our empirical results with other studies we see similar patterns qualitatively. As in [Hufe et al. \(2022b\)](#), [Brunori et al. \(2021\)](#), [Marrero and Rodríguez \(2012\)](#) or [Carranza \(2022\)](#) we see that Nordic countries and countries such as Germany or the Netherlands have low IOP, Southern countries have high IOP and there is heterogeneity in Eastern countries. [Hufe et al. \(2022b\)](#) takes a multidimensional approach and shows that this divide between north and south is reduced when considering also IOP in education and health.

We report more detailed results in [Table 2](#) where for each country we can see the income inequality, the plug-in and debiased estimates based on the best performing machine learner, relative IOP, the best performing machine learner, the cross-validated root mean squared error of the first step for the best performing machine learner and the sample size.

Country	Mean	Gini	Plug-in	Debiased	Debiased/Gini (%)	Best Performer	RMSE	n
Austria	29824	0.268	0.073	0.079 (0.066,0.091)	29 (25,33)	Ridge	16364	5119
Belgium	28432	0.237	0.099	0.092 (0.083,0.101)	39 (36,42)	CIF	12542	6359
Bulgaria	6603	0.415	0.238	0.244 (0.222,0.267)	59 (56,62)	CIF	8073	5906
Switzerland	52753	0.273	0.088	0.073 (0.056,0.09)	27 (21,32)	CIF	30580	4705
Cyprus	20281	0.299	0.101	0.096 (0.077,0.115)	32 (27,38)	Ridge	16083	4647
Czechia	12215	0.23	0.069	0.065 (0.056,0.073)	28 (25,32)	XGB	5944	6330
Germany	28813	0.271	0.073	0.061 (0.047,0.074)	22 (18,27)	XGB	17261	5967
Denmark	36768	0.26	0.074	0.013 (-0.009,0.034)	5 (-3,13)	CB	20807	2068
Estonia	14310	0.283	0.067	0.072 (0.063,0.081)	26 (22,29)	Lasso	7026	5649
Greece	9915	0.31	0.129	0.12 (0.109,0.131)	39 (36,42)	CB	6985	14144
Spain	17732	0.326	0.148	0.144 (0.136,0.152)	44 (42,46)	CIF	10587	16864
Finland	29305	0.263	0.033	0.035 (0.019,0.052)	13 (8,19)	Lasso	17033	4398
France	27014	0.274	0.105	0.094 (0.082,0.106)	34 (31,38)	CIF	15801	7821
Croatia	8916	0.281	0.11	0.1 (0.091,0.109)	36 (33,38)	CIF	4625	7120
Hungary	6929	0.282	0.094	0.082 (0.061,0.104)	29 (23,35)	CIF	4673	4557
Ireland	32223	0.278	0.135	0.107 (0.087,0.127)	39 (33,44)	CB	19362	3653
Italy	20004	0.319	0.125	0.125 (0.117,0.132)	39 (37,41)	CIF	12041	16359
Lithuania	10383	0.347	0.128	0.12 (0.1,0.139)	34 (30,39)	XGB	7150	3643
Luxembourg	45490	0.329	0.145	0.124 (0.1,0.147)	38 (32,43)	CIF	35653	3516
Latvia	10789	0.337	0.105	0.105 (0.09,0.12)	31 (27,35)	Lasso	7129	3337
Malta	19271	0.266	0.104	0.101 (0.088,0.114)	38 (34,42)	CIF	9280	3623
Netherlands	30636	0.259	0.069	0.046 (0.017,0.075)	18 (8,28)	CIF	18445	4441
Norway	41855	0.245	0.066	0.058 (0.04,0.077)	24 (17,31)	XGB	21998	2364
Poland	8621	0.292	0.118	0.11 (0.103,0.118)	38 (36,40)	CIF	5014	14281
Portugal	12032	0.305	0.144	0.13 (0.121,0.138)	42 (40,45)	XGB	7217	13776
Romania	4826	0.343	0.186	0.197 (0.185,0.209)	57 (55,60)	CIF	2667	5932
Serbia	3984	0.329	0.146	0.138 (0.126,0.149)	42 (39,45)	CIF	2448	5648
Sweden	29189	0.305	0.101	0.095 (0.057,0.132)	31 (21,41)	XGB	28923	1894
Slovakia	9197	0.221	0.076	0.072 (0.064,0.079)	32 (29,36)	CIF	3483	5727

Table 2: EUSILC Gini IOP: debiased vs. plug-in estimators, bestmachine learner and sample size

The bias of the plug-in estimator varies from country to country. Denmark, Croatia, Luxembourg, France, Poland, Ireland, Portugal and Romania have plug-in estimates outside the 95 % confidence interval of the debiased estimates. In Figure 5 we see the difference between the (relative) debiased and plug-in estimators. For most countries the plug-in estimator overestimates IOp but this is not always the case. The differences can be quite stark, in Denmark we have almost 25 percentage point difference. In the rest of the countries the differences are much smaller but we still observe differences of 5-10 percentage points in some countries.

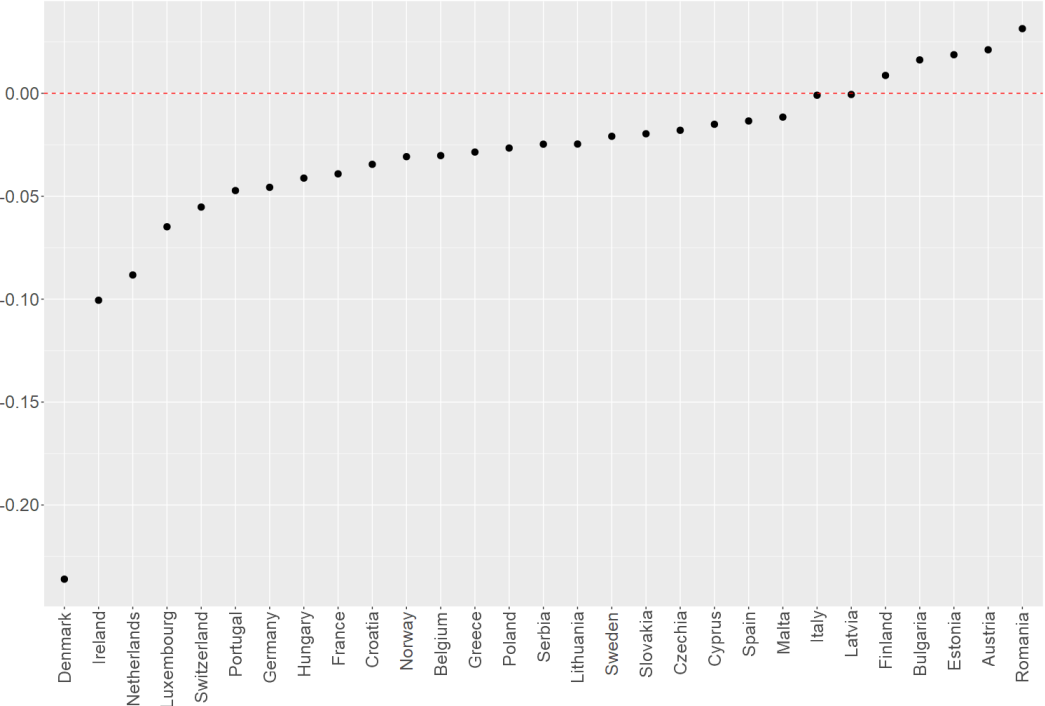


Figure 5: Difference between relative debiased IOp and plug-in IOp.

As already pointed out, the debiased estimates are much less sensitive to the choice of the machine learner. In Figure 6 we report the debiased and plug-in estimates for the 6 different machine learners. We can see that the debiased estimates are much more concentrated and near to each other than the plug-in estimates. The plug-in estimates are much more dispersed and in some cases there are differences of 60 percentage points from one machine learner to the other. Even differences of 30-40 percentage points between plug-in estimates using different machine learners are not uncommon. This result alone shows the importance of using locally robust debiased estimators when estimating IOp.

Finally, we want to have an idea of which circumstances are the most relevant to explain inequality. To do this, we compute the relative change in IOp when we exclude a given circumstance. In Table 3 we report the circumstance in each country for which this relative

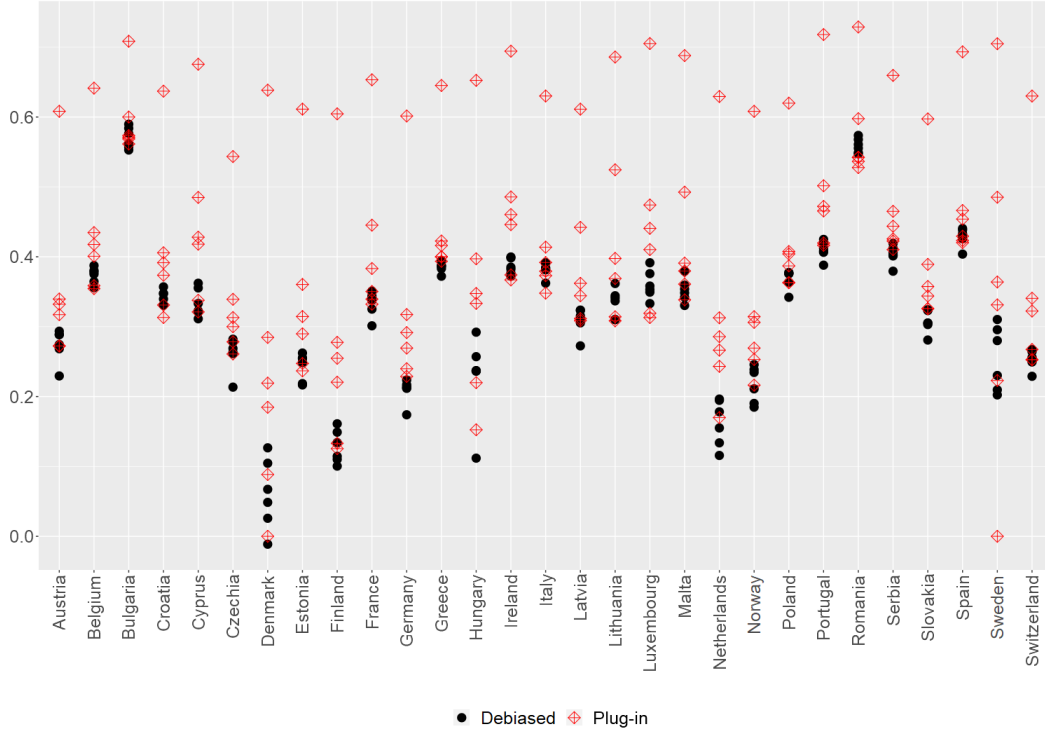


Figure 6: Debiased vs plug-in estimates for the 6 different machine learners

change is the greatest. For instance, in Austria the most important circumstance is living with the father. This means that if we exclude this circumstance, the relative IOp would decrease 2.6%. We see that the most important circumstance varies from country to country. However, parental education and occupation stand out as recurrent circumstances which help explain inequality the most. For instance in countries such as the Netherlands or Germany, excluding parental occupation leads to a 12-13% decrease in IOp. Meanwhile, excluding maternal education leads to a 4% decrease of IOp in Czechia or almost a 13% decrease in IOp in Hungary. This is in line with recent findings in [Hufe et al. \(2022b\)](#) who attribute the increase in IOp in the US mainly to the increased importance of parental occupation and education. Here we go a bit deeper, as we are able to separate paternal and maternal influences and provide uncertainty quantification. We see that while for the mother it is the education which matters the most to explain inequality, for the father it is the occupation which matters the most.

It might seem surprising that sex does not appear as an important contributor to explaining inequality in any country. However, as pointed out in [Hufe et al. \(2022a\)](#), this is expected since we are looking at equalised household income. Hence, any intra-household differences are not accounted for. Any explanatory power sex might have in our analysis would come only from single-headed households.

Country	Most important circumstance	Relative importance	CI
Austria	Living with the father	0.026	(0.022, 0.03)
Belgium	Father's education	0.038	(0.036, 0.04)
Bulgaria	Father's occupation	0.031	(0.027, 0.035)
Switzerland	Financial situation	0.051	(0.049, 0.053)
Cyprus	Population	0.017	(0.011, 0.023)
Czechia	Mother's education	0.044	(0.042, 0.046)
Germany	Father's occupation	0.131	(0.127, 0.135)
Denmark	Number of adults	0.689	(0.681, 0.697)
Estonia	Mother's education	0.064	(0.06, 0.068)
Greece	Country of birth	0.029	(0.027, 0.031)
Spain	Father's occupation	0.046	(0.044, 0.048)
Finland	Country of birth	0.057	(0.049, 0.065)
France	Father's occupation	0.089	(0.085, 0.093)
Croatia	Mother's education	0.033	(0.031, 0.035)
Hungary	Mother's education	0.126	(0.124, 0.128)
Ireland	Holidays	0.058	(0.054, 0.062)
Italy	Father's occupation	0.022	(0.02, 0.024)
Lithuania	Population	0.098	(0.092, 0.104)
Luxembourg	Country of birth	0.079	(0.077, 0.081)
Latvia	Number of adults	0.026	(0.022, 0.03)
Malta	Father's managerial status	0.075	(0.073, 0.077)
Netherlands	Father's occupation	0.12	(0.116, 0.124)
Norway	Father's managerial status	0.21	(0.202, 0.218)
Poland	Father's occupation	0.038	(0.036, 0.04)
Portugal	Father's occupation	0.052	(0.048, 0.056)
Romania	Population	0.037	(0.033, 0.041)
Serbia	Holidays	0.072	(0.068, 0.076)
Sweden	Financial situation	0.255	(0.249, 0.261)
Slovakia	Mother's education	0.053	(0.051, 0.055)

Table 3: Relative importance of most important circumstance in each country. The relative importance of a circumstance is computed by estimating the relative effect of dropping that circumstance.

6 Appendices

A Proofs of General Results

A.1 Main results

Proof of Lemma 1: Note that

$$\mathbb{E}[\delta(W_i, W_j, \gamma_0)(c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j))] = \mathbb{E}[\alpha_0(W_i, W_j, \gamma_0)(c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j))],$$

because

$$\mathbb{E}[(\delta(X_i, X_j, \gamma_0) - \alpha_0(X_i, X_j))(c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j))] = 0,$$

$\alpha_0(X_i, X_j) = \Pi_{\mathcal{S}}(\delta(X_i, X_j, \gamma_0))$ and $c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j) \in \Gamma + \Gamma \subseteq \mathcal{S}$. Since $\alpha_0(\cdot, x) \in \Gamma$ and $\alpha_0(x, \cdot) \in \Gamma$ for all $x \in \mathcal{X}$,

$$\mathbb{E}_{F_\tau}[\alpha_0(X_i, X_j)(c_1(Y_i - \gamma_\tau(X_i)) + c_2(Y_j - \gamma_\tau(X_j)))] = 0.$$

Thus, by the chain rule and because $\alpha_0 \in \mathcal{S}$,

$$\begin{aligned} \frac{d}{d\tau} \mathbb{E}[g(W_i, W_j, \gamma(F_\tau), \theta)] &= \frac{d}{d\tau} \mathbb{E}[\delta(X_i, X_j, \gamma_0)(c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j))] \\ &= \frac{d}{d\tau} \mathbb{E}[\alpha_0(X_i, X_j)(c_1\gamma_\tau(X_i) + c_2\gamma_\tau(X_j))] \\ &= \frac{d}{d\tau} \mathbb{E}_{F_\tau}[\alpha_0(X_i, X_j)(c_1Y_i + c_2Y_j - c_1\gamma_0(X_i) - c_2\gamma_0(X_j))] \\ &= \int \int \phi(w_i, w_j, \gamma_0, \alpha_0, \theta) K_H(dw_i, dw_j) \end{aligned}$$

with $\phi(w_i, w_j, \gamma_0, \alpha_0, \theta) = \alpha_0(x_i, x_j)(c_1y_i + c_2y_j - c_1\gamma_0(x_i) - c_2\gamma_0(x_j))$, which gives the desired result. ■

Proof of Lemma 2: The proof of (i) follows from $\alpha = \delta$. For (ii), use the short notation $\delta_{ij}(\gamma) \equiv \delta(X_i, X_j, \gamma)$, and note that by iterated expectations and independence, we can write

$$\begin{aligned} \delta_{ij}(\gamma) &= \mathbb{E}[\delta_{ij}(\gamma)|X_i] + \mathbb{E}[\delta_{ij}(\gamma)|X_j] - \mathbb{E}[\delta_{ij}(\gamma)] + U_{ij}, \\ &\equiv \tilde{\delta}_{ij}(\gamma) + U_{ij}, \end{aligned}$$

where $\mathbb{E}[U_{ij}|X_i] = \mathbb{E}[U_{ij}|X_j] = 0$. This expansion implies that only $\tilde{\delta}_{ij}$ matters for the derivative in (3.2), so δ_{ij} could be replaced everywhere by $\tilde{\delta}_{ij}$. Thus, substituting δ_{ij} from

the last display in the expression for α_0 we obtain

$$\begin{aligned}\alpha_0(X_i, X_j) &= \Pi_{\mathcal{S}}(\delta(X_i, X_j, \gamma_0)) \\ &= \Pi_{\mathcal{S}}\mathbb{E}[\delta_{ij}(\gamma_0)|X_i] + \Pi_{\mathcal{S}}\mathbb{E}[\delta_{ij}(\gamma_0)|X_j] - \mathbb{E}[\delta_{ij}(\gamma)] \\ &= \Pi_{\Gamma}\mathbb{E}[\delta_{ij}(\gamma_0)|X_i] + \Pi_{\Gamma}\mathbb{E}[\delta_{ij}(\gamma_0)|X_j] - \mathbb{E}[\delta_{ij}(\gamma)],\end{aligned}$$

where last expression uses that $\mathcal{S} = \Gamma + \Gamma$. ■

Proof of Lemma 3: First of all, note that the U-FSIF in (3.8) is not necessarily symmetric. However, without loss of generality we can consider the symmetric version

$$\phi^*(w_i, w_j, \gamma_0, \alpha_0) = \frac{1}{2}(\phi(w_i, w_j, \gamma_0, \alpha_0) + \phi(w_j, w_i, \gamma_0, \alpha_0)),$$

where ϕ is as in (3.8). We can decompose ϕ^* as

$$\phi^*(W_i, W_j, \gamma, \alpha) = \phi_1^*(W_i, \gamma, \alpha) + \phi_1^*(W_j, \gamma, \alpha) + \xi^*(W_i, W_j, \gamma, \alpha),$$

where $\xi^*(w_i, w_j, \gamma, \alpha) = \phi^*(w_i, w_j, \gamma, \alpha) - \phi_1^*(w_i, \gamma, \alpha) - \phi_1^*(w_j, \gamma, \alpha)$ is a degenerate kernel. By integrating with respect to K_H it follows that the FSIF is given by $2\phi_1^*(w, \gamma_0, \alpha_0)$. Noting that $2\phi_1^*(w, \gamma_0, \alpha_0) = \phi_1(w, \gamma_0, \alpha_0) + \phi_2(w, \gamma_0, \alpha_0)$ and that

$$\phi_1(w, \gamma_0, \alpha_0) = c_1\alpha_{01}(x)(y - \gamma_0(x)), \quad \phi_2(w, \gamma_0, \alpha_0) = c_2\alpha_{02}(x)(y - \gamma_0(x)),$$

we get the desired result. ■

B Proofs of Inequality of Opportunity

Proof of Proposition 1: We want to compute

$$\frac{\partial}{\partial \tau}\mathbb{E}(g(W_i, W_j, \gamma(F_\tau), \theta)) = \frac{\partial}{\partial \tau}\mathbb{E}[\theta(\gamma_\tau(X_i) + \gamma_\tau(X_j))] - \frac{\partial}{\partial \tau}\mathbb{E}(|\gamma_\tau(X_i) - \gamma_\tau(X_j)|).$$

The first term is already in the form of (3.2) so we can directly apply Lemma 1 to find

$$\phi_1(w_i, w_j, \gamma_0, \theta) = \theta(y_i + y_j - \gamma_0(x_i) - \gamma_0(x_j)).$$

Define short notation $\Delta_0 \equiv \gamma_0(X_i) - \gamma_0(X_j)$. First we find the Gateaux derivative of the map $\Delta \mapsto \mathbb{E}(|\Delta|)$ at Δ_0 for directions $v \in V = \{v : v(x_{ij}) = \tilde{v}(x_i) - \tilde{v}(x_j), \tilde{v} \in \Gamma\}$. We use that $\mathbb{E}(|\Delta|) = \mathbb{E}([1(\Delta \geq 0) - 1(\Delta < 0)]\Delta)$. Consider a deviation in the direction of the function v , $\Delta_0 + tv$, where, under Assumption 1(i) v is such that $\|v\|_\infty = \sup_{x_i, x_j \in \mathcal{X}^2} |v(x_i, x_j)| \leq C$

and $t > 0$, then

$$\begin{aligned} & \left[\mathbb{E} \left([1(\Delta_0 \geq -tv) - 1(\Delta_0 < -tv)][\Delta_0 + tv] \right) - \mathbb{E}(|\Delta_0|) \right] / t = \\ & \underbrace{\left[\mathbb{E} \left([1(\Delta_0 \geq -tv) - 1(\Delta_0 < -tv)]\Delta_0 \right) - \mathbb{E}(|\Delta_0|) \right] / t}_{(1)} + \underbrace{\mathbb{E} \left([1(\Delta_0 \geq -tv) - 1(\Delta_0 < -tv)]v \right)}_{(2)}. \end{aligned}$$

Note that

$$\left| (2) - \mathbb{E} \left([1(\Delta_0 > 0) - 1(\Delta_0 < 0)]v \right) \right| \leq 2C \mathbb{E}[1(-tC \leq \Delta_0 \leq tC)],$$

so (2) $\rightarrow \mathbb{E} \left([1(\Delta_0 > 0) - 1(\Delta_0 < 0)]v \right)$ by Assumption 1(ii) as $t \downarrow 0$. Alternatively, under 1(iii)

$$\left| (2) - \mathbb{E} \left([1(\Delta_0 > 0) - 1(\Delta_0 < 0)]v \right) \right| \leq 2C \mathbb{E}[1(-tC \leq \Delta_0 \leq tC)1(X_i \neq X_j)] \rightarrow 0.$$

Now we show that (1) $\rightarrow 0$ as $t \downarrow 0$. We use that $\Delta_0 = \Delta_0^+ - \Delta_0^-$, where $\Delta_0^+ = 1(\Delta_0 \geq 0)\Delta_0$ and $\Delta_0^- = -1(\Delta_0 < 0)\Delta_0$.

$$\begin{aligned} |(1)| &= \left| \left[\mathbb{E}([1 - 2 \cdot 1(\Delta_0 < -tv)]\Delta_0^+ - [-1 + 2 \cdot 1(\Delta_0 < -tv)]\Delta_0^-) - \mathbb{E}(|\Delta_0|) \right] / t \right| \\ &= \left| \left[-2\mathbb{E}[1(\Delta_0 < -tv)\Delta_0^+] - 2\mathbb{E}[1(\Delta_0 \geq -tv)\Delta_0^-] + \underbrace{\mathbb{E}[\Delta_0^+ + \Delta_0^- - |\Delta_0|]}_{=0} \right] / t \right| \\ &\leq 2 \left| \left[\mathbb{E}(1(\Delta_0 < -tv)1(\Delta_0 \geq 0)\Delta_0) \right] / t \right| + 2 \left| \left[\mathbb{E}(1(\Delta_0 \geq -tv)1(\Delta_0 < 0)\Delta_0) \right] / t \right| \quad (\star). \end{aligned}$$

In the first equality we use that $\Delta_0 = \Delta_0^+ - \Delta_0^-$ and that $[1(\Delta_0 \geq -tv) - 1(\Delta_0 < -tv)] = 1 - 2 \cdot 1(\Delta_0 < -tv) = -1 + 2 \cdot 1(\Delta_0 \geq -tv)$. In the second equality we note that $\Delta_0^+ + \Delta_0^- = [1(\Delta_0 \geq 0) - 1(\Delta_0 < 0)]\Delta_0 = |\Delta_0|$. In the inequality we use the triangle inequality.

Now we note that $1(\Delta_0 < -tv) \leq 1(\Delta_0 \leq tC)$ and $1(\Delta_0 \geq -tv) \leq 1(\Delta_0 \geq -tC)$, then

$$\begin{aligned} (\star) &\leq 2 \left| \mathbb{E}[1(\Delta_0 \leq tC)1(\Delta_0 \geq 0)\Delta_0] / t \right| + 2 \left| \mathbb{E}[1(\Delta_0 \geq -tC)1(\Delta_0 < 0)\Delta_0] / t \right| \\ &= 2 \left| \mathbb{E}[1(0 \leq \Delta_0 \leq tC)\Delta_0] / t \right| + 2 \left| \mathbb{E}[1(-tC \leq \Delta_0 < 0)\Delta_0] / t \right|. \end{aligned}$$

Hence, by Assumption 1(ii) or (iii), $(\star) \rightarrow 0$ and the Gateaux derivative in direction v of

$\Delta \mapsto \mathbb{E}(|\Delta|)$ is

$$\mathbb{E}\left([1(\Delta_0 > 0) - 1(\Delta_0 < -0)]v\right).$$

Then,

$$\frac{\partial \mathbb{E}(|\gamma_\tau(X_i) - \gamma_\tau(X_j)|)}{\partial \tau} = \frac{\partial}{\partial \tau} \mathbb{E}[\delta(X_i, X_j, \gamma_0)(\gamma_\tau(X_i) - \gamma_\tau(X_j))],$$

where $\delta(X_i, X_j, \gamma) = \text{sgn}(\gamma(X_i) - \gamma(X_j))$. By Lemma 1 the U-FSIF of the second term is

$$\phi_2(w_i, w_j, \gamma_0) = \alpha_0(x_i, x_j, \gamma_0)(y_i - y_j - \gamma_0(x_i) + \gamma_0(x_j))$$

where $\alpha_0(x_i, x_j, \gamma) = \Pi_S(\delta(x_i, x_j, \gamma))$. Thus, the U-FSIF is

$$\phi(w_i, w_j, \gamma_0, \alpha_0, \theta) = \phi_1(w_i, w_j, \gamma_0, \theta) - \phi_2(w_i, w_j, \gamma_0, \alpha_0).$$

■

Proof of Proposition 2: Let $\hat{\beta}_{ij} = \hat{\Delta}_{\hat{\gamma}_l} - \Delta_0$ and $\beta_n = \sup_{i,j} |\beta_{ij}|$. First of all we notice that

$$\begin{aligned} \hat{\alpha}_l - \alpha_0 &= 1(\hat{\Delta}_{\hat{\gamma}_l} > 0) - 1(\Delta_0 > 0) + 1(\Delta_0 < 0) - 1(\hat{\Delta}_{\hat{\gamma}_l} < 0) \\ &= 1(\Delta_0 > -\hat{\beta}_{ij}) - 1(\Delta_0 > 0) + 1(\Delta_0 < 0) - 1(\Delta_0 < -\hat{\beta}_{ij}) \\ &\leq 1(\Delta_0 > -\beta_n) - 1(\Delta_0 > 0) + 1(\Delta_0 < 0) - 1(\Delta_0 < -\beta_n) \\ &= 1(-\beta_n < \Delta_0 \leq 0) + 1(-\beta_n \leq \Delta_0 < 0) \\ &\leq 2 \cdot 1(-\beta_n \leq \Delta_0 \leq 0), \end{aligned}$$

where in the second equality we have added and subtracted Δ_0 inside the first and last indicators. Hence, for Assumption 2 (iii) we have that

$$\begin{aligned} \mathbb{E}(|\hat{\alpha}_l - \alpha_0|^2 | N_l^c) &\leq 4\mathbb{E}[1(-\beta_n \leq \Delta_0 \leq 0) | N_l^c] \\ &= 4(F_{\Delta_0}(0) - F_{\Delta_0}(-\beta_n)) \rightarrow_p 0, \end{aligned}$$

where convergence follows from continuity of F_{Δ_0} (implied by absolute continuity) and the

fact that $a_n = o_p(n^{-1/4})$ implies $\beta_n = o_p(n^{-1/4})$. For Assumption 2 (iv) we have that

$$\begin{aligned}
& \sqrt{n} \int \int (\hat{\alpha}_l(x_i, x_j) - \alpha_0(x_i, x_j))(\hat{\gamma}_l(x_i) - \gamma_0(x_i) - \hat{\gamma}_l(x_j) + \gamma_0(x_j)) F_0(dw_i) F_0(dw_j) \\
& \leq 2\sqrt{n} \int \int 1(-\beta_n \leq \Delta_0 \leq 0) \beta_n F_0(dw_i) F_0(dw_j) \\
& = 2\sqrt{n} (F_{\Delta_0}(0) - F_{\Delta_0}(-\beta_n)) \beta_n \\
& \leq 2C \sqrt{n} \beta_n^2 \\
& = 2C \sqrt{n} o_p(n^{-1/4}) o_p(n^{-1/4}) \\
& = o_p(1),
\end{aligned}$$

where the second inequality follows since F_{Δ_0} absolutely continuous with bounded density around zero implies that F_{Δ_0} is Lipschitz continuous at zero of some constant C by the mean value theorem. To use condition (ii) in the Proposition instead of condition (i) first note that

$$\begin{aligned}
\mathbb{E}(|\hat{\alpha}_l - \alpha_0|^2 | N_l^c) &= \mathbb{E}(|\text{sgn}(\Delta_{\hat{\gamma}_l}) - \text{sgn}(\Delta_0)|^2 | N_l^c) \\
&\leq 4\mathbb{P}(\text{sgn}(\Delta_{\hat{\gamma}_l}) \neq \text{sgn}(\Delta_0) | N_l^c) \\
&\leq 4\mathbb{P}(\text{sgn}(\Delta_{\hat{\gamma}_l}) \neq \text{sgn}(\Delta_0) | N_l^c, X_i \neq X_j) \\
&= 4\mathbb{P}(\Delta_{\hat{\gamma}_l} \Delta_0 \leq 0 | N_l^c, X_i \neq X_j).
\end{aligned}$$

By assumption $\mathbb{P}(|\Delta_0| \leq \eta) = 0$ (conditional on $X_i \neq X_j$) and

$$\begin{aligned}
\mathbb{P}(\Delta_{\hat{\gamma}_l} \Delta_0 \leq 0 | N_l^c, X_i \neq X_j) &= \mathbb{P}(\Delta_{\hat{\gamma}_l} \Delta_0 \leq 0, |\Delta_0| > \eta | N_l^c, X_i \neq X_j) \\
&\quad + \mathbb{P}(\Delta_{\hat{\gamma}_l} \Delta_0 \leq 0, |\Delta_0| \leq \eta | N_l^c, X_i \neq X_j) \\
&\leq \mathbb{P}(|\Delta_{\hat{\gamma}_l} - \Delta_0| > \eta | N_l^c, X_i \neq X_j) \\
&\leq \frac{\mathbb{E}(|\Delta_{\hat{\gamma}_l} - \Delta_0|^2 | N_l^c, X_i \neq X_j)}{\eta^2} \rightarrow_p 0.
\end{aligned}$$

The first inequality follows since $\{\Delta_{\hat{\gamma}_l} \Delta_0 \leq 0, |\Delta_0| > \eta\} \subseteq \{|\Delta_{\hat{\gamma}_l} - \Delta_0| > \eta\}$ and because the probability in the second line is bounded by $\mathbb{P}(|\Delta_0| \leq \eta) = 0$. The second inequality follows from the conditional Markov inequality and the convergence follows from mean square consistency of $\hat{\gamma}_l$. Hence, Assumption 2 (iii) follows. Finally, multiplying by \sqrt{n}

$$\begin{aligned}
\sqrt{n} \mathbb{E}(|\hat{\alpha}_l - \alpha_0|^2 | N_l^c, X_i \neq X_j) &\leq 4\sqrt{n} \mathbb{P}(|\Delta_{\hat{\gamma}_l} - \Delta_0| > \eta | N_l^c, X_i \neq X_j) \\
&\leq 4\sqrt{n} \frac{\mathbb{E}(|\Delta_{\hat{\gamma}_l} - \Delta_0|^2 | N_l^c, X_i \neq X_j)}{\eta^2} \rightarrow_p 0.
\end{aligned}$$

The convergence in probability follows from $\|\hat{\gamma}_l - \gamma_0\| = o_p(n^{-1/4})$. Hence we have also that $\|\hat{\alpha}_l - \alpha_0\| = o_p(n^{-1/4})$ and Assumption 2 (iv) follows. ■

Proof of Proposition 3: We need to show that the general assumptions of the asymptotic theory in the Online Appendix hold in this example. Assumption 1 (i) follows since using the C_r inequality multiple times and by the reverse triangle inequality

$$\int \int |g(w_i, w_j, \hat{\gamma}_l, \theta_0) - g(w_i, w_j, \gamma_0, \theta_0)|^2 F_0(dw_i) F_0(dw_j) \leq C \int |\hat{\gamma}_l(x) - \gamma_0(x)|^2 F_0(dx),$$

which converges to zero by Assumption 2 (ii). Similarly,

$$\begin{aligned} & \int \int |\phi(w_i, w_j, \hat{\gamma}_l, \alpha_0, \theta_0) - \phi(w_i, w_j, \gamma_0, \alpha_0, \theta_0)|^2 F_0(dw_i) F_0(dw_j) \\ &= \int \int |(\theta_0 - \alpha_0(x_i, x_j))(\gamma_0(x_i) - \hat{\gamma}_l(x_i) + \gamma_0(x_j) - \hat{\gamma}_l(x_j))|^2 F_0(dw_i) F_0(dw_j) \\ &\leq C \int |\hat{\gamma}_l(x) - \gamma_0(x)|^2 F_0(dw) \rightarrow_p 0, \end{aligned}$$

where we used the C_r inequality and boundedness of $\theta_0 - \alpha_0(x_i, x_j)$. Hence Assumption 1 (ii) is also satisfied. For Assumption 1 (iii)

$$\begin{aligned} & \int \int |\phi(w_i, w_j, \gamma_0, \hat{\alpha}_l, \theta_0) - \phi(w_i, w_j, \gamma_0, \alpha_0, \theta_0)|^2 F_0(dw_i) F_0(dw_j) \\ &= \int \int |(y_i - y_j - \gamma_0(x_i) + \gamma_0(x_j))(\alpha_0(x_i, x_j) - \hat{\alpha}_l(x_i, x_j))|^2 F_0(dw_i) F_0(dw_j) \\ &\leq C \int \int |\alpha_0(x_i, x_j) - \hat{\alpha}_l(x_i, x_j)|^2 F_0(dw) \rightarrow_p 0, \end{aligned}$$

where in the inequality we used Assumption 2 (i) and the convergence to zero follows from Assumption 2 (iii). To check Assumption 2 (i) note that

$$\begin{aligned} & \sqrt{n} \int \int \hat{\xi}(w_i, w_j) F_0(dw_i) F_0(dw_j) \\ &= \sqrt{n} \int \int (\hat{\alpha}_l(x_i, x_j) - \alpha_0(x_i, x_j)) (\hat{\gamma}_l(x_i) - \gamma_0(x_i) - \hat{\gamma}_l(x_j) + \gamma_0(x_j)) F_0(dw_i) F_0(dw_j) \rightarrow_p 0, \end{aligned}$$

where the convergence follows by Assumption 2 (iv). Also,

$$\begin{aligned} & \int \int |\hat{\xi}(w_i, w_j)|^2 F_0(dw_i) F_0(dw_j) \\ &= \int \int (\hat{\alpha}_l(x_i, x_j) - \alpha_0(x_i, x_j))^2 (\hat{\gamma}_l(x_i) - \gamma_0(x_i) - \hat{\gamma}_l(x_j) + \gamma_0(x_j))^2 F_0(dw_i) F_0(dw_j) \rightarrow_p 0, \end{aligned}$$

where the convergence follows from boundedness of α_0 and by Assumption 2 (ii). Assumption

3 (i) is global robustness and follows from visual inspection. Assumption 3 (ii) follows from Assumption 2 (ii) and by result 1.8.4 in Yamamuro (1974). To use the last result note that the second order Gateaux derivative of $\bar{\psi}(\gamma, \alpha_0, \theta_0)$ with respect to γ for $\|\gamma - \gamma_0\|$ small enough is 0 under Assumption 1 (ii) since then the sgn function has zero slope a.s. For the conditions for the consistency of the variance let $\mathbb{E}_{n,i} = (n-1)^{-1} \sum_{j \neq i}$ and note that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n |\hat{g}_{-i} - g_{-i}|^2 &\leq \frac{C}{n} \sum_{i=1}^n |\mathbb{E}_{n,i}[\hat{\gamma}(X_i) + \hat{\gamma}(X_j)]|^2 |\hat{\theta} - \theta_0|^2 \\ &\quad + \frac{C\theta_0}{n} \sum_{i=1}^n |\mathbb{E}_{n,i}[\hat{\gamma}(X_i) + \hat{\gamma}(X_j) - \gamma_0(X_i) - \gamma_0(X_j)]|^2 \\ &\quad + \frac{C}{n} \sum_{i=1}^n |\mathbb{E}_{n,i}[\gamma_0(X_i) - \gamma_0(X_j) - \hat{\gamma}(X_i) + \hat{\gamma}(X_j)]|^2. \end{aligned}$$

Now let us show that each term vanishes. First we note that $(1/n) \sum_{i=1}^n |\mathbb{E}_{n,i}[\hat{\gamma}(X_i) + \hat{\gamma}(X_j)]|^2 = O_p(1)$, let $C_\varepsilon > 0$, then

$$\begin{aligned} &\mathbb{P} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{(n-1)^2} \left(\sum_{j \neq i} \hat{\gamma}(X_i) + \hat{\gamma}(X_j) \right)^2 > C_\varepsilon \right] \\ &\leq \frac{\mathbb{E} \left[\frac{1}{(n-1)^2} \left((n-1)\hat{\gamma}(X_i) + \sum_{j \neq i} \hat{\gamma}(X_j) \right)^2 \right]}{C_\varepsilon} \\ &\leq \frac{C \left(\frac{1}{n-1} \mathbb{E}[(\hat{\gamma}(X_i) - \gamma_0(X_i) + \gamma_0(X_i))^2] + \frac{1}{n-1} \mathbb{E} \left[\left(\sum_{j \neq i} \hat{\gamma}(X_j) \right)^2 \right] \right)}{C_\varepsilon} \\ &\leq \frac{C \left(\frac{1}{n-1} C(\mathbb{E}[(\hat{\gamma}(X_i) - \gamma_0(X_i))^2] + \mathbb{E}[\gamma_0(X_i)^2]) + C(\mathbb{E}[(\hat{\gamma}(X_j) - \gamma_0(X_j))^2] + \mathbb{E}[\gamma_0(X_j)^2]) \right)}{C_\varepsilon} \\ &\rightarrow_p \frac{C\mathbb{E}[\gamma_0(X_j)^2]}{C_\varepsilon}. \end{aligned}$$

The first inequality uses Markov's inequality, the second one uses C_r inequality, the third one uses C_r and Cauchy Schwartz inequalities and the convergence follows from L2 convergence of $\hat{\gamma}$. Hence, for all $\varepsilon > 0$ and sufficiently large n we can choose $C_\varepsilon > 0$ sufficiently large so as to make $\mathbb{P} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{(n-1)^2} \left(\sum_{j \neq i} \hat{\gamma}(X_i) + \hat{\gamma}(X_j) \right)^2 > C_\varepsilon \right] < \varepsilon$. Hence the first term is $O_p(1)|\hat{\theta} - \theta_0|^2 \rightarrow_p 0$ by consistency of $\hat{\theta}$. Following the same steps we can bound the probability that the second terms exceed some $\varepsilon > 0$ by $\mathbb{E}[(\hat{\gamma} - \gamma_0)^2] \rightarrow_p 0$ so the second

term vanishes. The exact same arguments cause the third term to vanish too. Now

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n |\hat{\phi}_{-i} - \phi_{-i}|^2 &\leq \frac{C}{n} \sum_{i=1}^n |\mathbb{E}_{n,i}[Y_i + Y_j - \hat{\gamma}(X_i) - \hat{\gamma}(X_j)]|^2 |\hat{\theta} - \theta_0|^2 \\
&+ \frac{C\theta_0}{n} \sum_{i=1}^n |\mathbb{E}_{n,i}[\gamma_0(X_i) + \gamma_0(X_j) - \hat{\gamma}(X_i) - \hat{\gamma}(X_j)]|^2 \\
&+ \frac{C}{n} \sum_{i=1}^n |\mathbb{E}_{n,i}[(\alpha_0(X_i, X_j) - \hat{\alpha}(X_i, X_j))(Y_i - Y_j - \hat{\gamma}(X_i) + \hat{\gamma}(X_j))]|^2 \\
&- \frac{C}{n} \sum_{i=1}^n |\mathbb{E}_{n,i}[\alpha_0(X_i, X_j)(Y_i - Y_j - \hat{\gamma}(X_i) + \hat{\gamma}(X_j))]|^2.
\end{aligned}$$

For the first term note that by the same arguments as before

$$\begin{aligned}
&\mathbb{P}\left(\frac{C}{n} \sum_{i=1}^n |\mathbb{E}_{n,i}[Y_i + Y_j - \hat{\gamma}(X_i) - \hat{\gamma}(X_j)]|^2 > C_\varepsilon\right) \\
&\leq \frac{\mathbb{E}\left[\frac{1}{(n-1)^2} \left(\sum_{j \neq i} Y_i + Y_j - \gamma_0(X_i) - \gamma_0(X_j) + \gamma_0(X_i) - \gamma_0(X_j) - \hat{\gamma}(X_i) - \hat{\gamma}(X_j)\right)^2\right]}{\varepsilon} \\
&\rightarrow_p C\mathbb{E}[(Y_j - \gamma_0(X_j))^2]/C_\varepsilon.
\end{aligned}$$

So the first term is also $O_p(1)|\hat{\theta} - \theta_0|^2 \rightarrow_p 0$. The second term vanishes as we have already shown. For the third term note that

$$\begin{aligned}
&\mathbb{P}\left(\frac{C}{n(n-1)^2} \sum_{i=1}^n \left| \sum_{j \neq i} [(\alpha_0(X_i, X_j) - \hat{\alpha}(X_i, X_j))(Y_i - Y_j - \Delta_0 - \Delta_{\hat{\gamma}})] \right|^2 > \varepsilon\right) \\
&\leq \frac{\mathbb{E}\left[\frac{C}{(n-1)^2} \left(\sum_{j \neq i} (\alpha_0(X_i, X_j) - \hat{\alpha}(X_i, X_j))(Y_i - Y_j - \Delta_0 + \Delta_0 - \Delta_{\hat{\gamma}})\right)^2\right]}{\varepsilon} \\
&\leq C \frac{\mathbb{E}\left[\frac{C}{(n-1)^2} \left(\sum_{j \neq i} (\alpha_0(X_i, X_j) - \hat{\alpha}(X_i, X_j))(Y_i - Y_j - \Delta_0)\right)^2\right]}{\varepsilon} \\
&+ C \frac{\mathbb{E}\left[\frac{C}{(n-1)^2} \left(\sum_{j \neq i} (\alpha_0(X_i, X_j) - \hat{\alpha}(X_i, X_j))(\Delta_0 - \Delta_{\hat{\gamma}})\right)^2\right]}{\varepsilon}.
\end{aligned}$$

Where we have used Markov's inequality and C_r inequality. For the first term above use Cauchy Schwartz, law of iterated expectations on (X_i, X_j) and a.s. finiteness of $\mathbb{E}[(Y_i - Y_j - \Delta_0)^2 | X_i, X_j]$ to show that it converges in probability to zero by L2 convergence of $\hat{\alpha}$. For the second term above use the fact that $(\alpha_0(X_i, X_j) - \hat{\alpha}(X_i, X_j))$ is bounded and L2 convergence

of $\hat{\gamma}$. Finally, for the fourth term use Markov's inequality as usual, add and subtract Δ_0 and use that

$$\mathbb{E}[\alpha_0(X_i, X_j)(Y_i - Y_j - \Delta_0)] = 0,$$

by global robustness and that $\alpha_0(X_i, X_j) \leq 1$ and L2 convergence of $\hat{\gamma}$ as before. Convergence of the Jacobian follows trivially since $\frac{\partial}{\partial \theta} \psi(w_i, w_j, \gamma, \alpha, \theta) = y_i + y_j$.

■

Online Appendix

1 On Uniqueness of the U-FSIF and the FSIF

The U-FSIF is unique up to the addition of functions with mean zero with respect to K_H for any H . This follows since, for any function $\xi(w_i, w_j)$ such that $\int \int \xi(w_i, w_j) K_H(dw_i, dw_j) = 0$,

$$\int \int [\phi(w_i, w_j, \gamma, \alpha) + \xi(w_i, w_j)] K_H(dw_i, dw_j) = \int \int \phi(w_i, w_j, \gamma, \alpha) K_H(dw_i, dw_j), \quad (1.1)$$

i.e. if ϕ is a U-FSIF so is $\phi + \xi$. If no restrictions are placed on the alternative distribution H , such functions $\xi(w_i, w_j)$ are called degenerate kernels (see, e.g., [Lee \(2019\)](#)).

For any symmetric U-FSIF ϕ , we define $\phi_1(W_i, \gamma, \alpha) = \int \phi(W_i, w, \gamma, \alpha) F_0(dw)$. Then, we can decompose any such $\phi(W_i, W_j, \gamma, \alpha)$ with mean zero as

$$\phi(W_i, W_j, \gamma, \alpha) = \phi_1(W_i, \gamma, \alpha) + \phi_1(W_j, \gamma, \alpha) + \xi(W_i, W_j, \gamma, \alpha),$$

where $\xi(w_i, w_j, \gamma, \alpha) = \phi(w_i, w_j, \gamma, \alpha) - \phi_1(w_i, \gamma, \alpha) - \phi_1(w_j, \gamma, \alpha)$ is a degenerate kernel. Hence,

$$\int \int \phi(w_i, w_j, \gamma, \alpha) K_H(dw_i, dw_j) = \int 2\phi_1(w, \gamma, \alpha) H(dw), \quad (1.2)$$

where the right hand side (RHS) coincides with the characterization of $2\phi_1$ as the FSIF in [Chernozhukov et al. \(2022\)](#). Thus, (1.1) shows that U-FSIFs are unique only up to the addition of degenerate kernels, while (1.2) shows all U-FSIFs give rise to the same FSIF (hence same first-order asymptotics).

EXAMPLE 1 (Variance of fitted values, cont.): ϕ_1 is

$$\phi_1(W_i, \gamma, \alpha) = [\gamma_0(X_i) - \mathbb{E}[Y_i]] \times (Y_i - \gamma_0(X_i)). \quad (1.3)$$

and

$$\xi(w_i, w_j) = (\mathbb{E}[Y_j] - \gamma_0(x_j)) (y_i - \gamma_0(x_i)) - (\mathbb{E}[Y_i] - \gamma_0(x_i)) (y_j - \gamma_0(x_j)), \quad (1.4)$$

which is a degenerate kernel. ■

An alternative approach based only on linear moments and FSIFs is to let $\bar{g}(w_i, \gamma_0, F_0, \theta) \equiv \int g(w_i, w_j, \gamma_0, \theta) F_0(dw_j)$ be the identifying moment function with unknown nuisance parameters (γ_0, F_0) . The derivative with respect to the first steps in this case is by the chain

rule

$$\frac{d}{d\tau}\mathbb{E}[\bar{g}(W_i, \gamma(F_\tau), F_\tau, \theta)] = \frac{d}{d\tau}\mathbb{E}[g(W_i, W_j, \gamma(F_\tau), \theta)] + \frac{d}{d\tau}\mathbb{E}[\bar{g}(W_i, \gamma(F_0), F_\tau, \theta)].$$

The first term of the RHS can be characterized by Equation (2.2) in the main text and the second term implies additional nuisance parameters. Hence, this alternative approach does not lead to a simpler methodology or simpler estimators than ours and it relies on our derivations for its completion. Thus, we recommend following instead our methods based on U-FSIFs.

The discussion about uniqueness also helps in constructing a U-FSIF from the FSIF: symmetrizing $2\phi_1$, i.e. by computing $\phi_1^*(w_i, w_j, \gamma, \alpha) = \phi_1(w_i, \gamma, \alpha) + \phi_1(w_j, \gamma, \alpha)$. However, we have found in our leading application below that an alternative construction of U-FSIF that we provide in the next section leads to simpler estimators such as our debiased IOP estimator. We give this general construction of U-FSIF for first steps that solve orthogonality restrictions and illustrate calculations in our running example and other examples below.

REMARK 1 (No first step estimation effect): A condition for the first step not having an effect on the first order asymptotics for the parameter of interest is $\phi_1(\cdot, \gamma_0, \alpha_0) \equiv 0$, i.e. that the U-FSIF is degenerate. We have shown that this condition does not hold in a number of applications, thereby invalidating inferences that do not account for first steps in standard errors. ■

EXAMPLE 1 (Variance of fitted values, cont.): $\phi_1 = 0$ only when γ_0 is constant. Thus, only when $\gamma_0(\cdot)$ is a constant there is no estimation effect from the first steps. When the fitted value is not constant, e.g. under IOP, there is an estimation effect from the first step, inference not accounting for the first step is invalid and valid inference on the variance of fitted values can be based on orthogonal quadratic moments proposed here. ■

2 Asymptotic theory

The aim of using a debiased moment function and cross-fitting is to be able to perform valid inference. First, we will show the key result that

$$\sqrt{n}\binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \psi(W_i, W_j, \hat{\gamma}_l, \hat{\alpha}_l, \theta_0) = \sqrt{n}\binom{n}{2}^{-1} \sum_{i < j} \psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0) + o_p(1). \quad (2.1)$$

This and other asymptotic results of this section are shown for generic first steps, not necessarily first steps satisfying orthogonality restrictions (e.g. machine learning estimates of

a density). The estimators $\hat{\alpha}_l$ for $\alpha_0(X_i, X_j) \equiv \Pi_{\mathcal{S}}(\delta(X_i, X_j, \gamma_0))$ depend on the choice of \mathcal{S} and the expression for $\delta(X_i, X_j, \gamma_0)$. In all our examples δ is known up to the first step γ_0 , so in the joint nonparametric case $\alpha_0 = \delta(\cdot, \gamma_0)$ is also known up to γ_0 and we can set $\hat{\alpha}_l = \delta(\cdot, \hat{\gamma}_l)$. For other cases we can obtain $\hat{\alpha}_l$ from Lemmas 1 and 2 in the main text by any machine learner method that estimates the projection of $\delta(\cdot, \hat{\gamma}_l)$ onto \mathcal{S} or Γ (such as Lasso, neural nets, sieves, etc.). For example, using Lemma 2 (ii) in the main text and letting $n_l = \sum_{j:(i,j) \notin I_l} 1$, we can first compute

$$\tilde{\alpha}_{1l}(X_i) = \frac{1}{n_l} \sum_{j \neq i} \delta_{ij}(\hat{\gamma}_l) \text{ and } \tilde{\alpha}_{2l}(X_j) = \frac{1}{n_l} \sum_{i \neq j} \delta_{ij}(\hat{\gamma}_l),$$

and then we can estimate $\alpha_{0r}(x)$, $r = 1, 2$, by the orthogonal projection of $\tilde{\alpha}_{rl}(X_i)$ onto Γ with any machine learning estimator and with observations not in I_l . Let now $|\cdot|$ and $\|\cdot\|$ be the Euclidean and L_2 norms, respectively. The terms \rightarrow_p and \rightarrow_d denote convergence in probability and distribution, respectively.

Assumption 4 $\mathbb{E}[|\psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)|^2] < \infty$ and

$$(i) \int \int |g(w_i, w_j, \hat{\gamma}_l, \theta_0) - g(w_i, w_j, \gamma_0, \theta_0)|^2 F_0(dw_i) F_0(dw_j) \rightarrow_p 0;$$

$$(ii) \int \int |\phi(w_i, w_j, \hat{\gamma}_l, \alpha_0, \theta_0) - \phi(w_i, w_j, \gamma_0, \alpha_0, \theta_0)|^2 F_0(dw_i) F_0(dw_j) \rightarrow_p 0;$$

$$(iii) \int \int |\phi(w_i, w_j, \gamma_0, \hat{\alpha}_l, \theta_0) - \phi(w_i, w_j, \gamma_0, \alpha_0, \theta_0)|^2 F_0(dw_i) F_0(dw_j) \rightarrow_p 0.$$

These are mild mean-square consistency conditions for $\hat{\gamma}_l$ and $\hat{\alpha}_l$ separately. A linearization argument like Equation (3.2) in the main text often implies that the left hand side of Assumption 4(i)-(ii) are bounded above by a constant times $\|\hat{\gamma}_l - \gamma_0\|^2$, so L_2 consistency suffices. Assumption 4(iii) typically follows from L_2 consistency of $\hat{\alpha}_l$, as for ϕ in Equation (3.3) in the main text. Define also the following interaction term

$$\hat{\xi}_l(w_i, w_j) = \phi(w_i, w_j, \hat{\gamma}_l, \hat{\alpha}_l, \theta_0) - \phi(w_i, w_j, \gamma_0, \hat{\alpha}_l, \theta_0) - \phi(w_i, w_j, \hat{\gamma}_l, \alpha_0, \theta_0) + \phi(w_i, w_j, \gamma_0, \alpha_0, \theta_0).$$

Assumption 5 For each $l = 1, \dots, L$, either i)

$$\sqrt{n} \int \int \hat{\xi}_l(w_i, w_j) F_0(dw_i) F_0(dw_j) \rightarrow_p 0, \int \int |\hat{\xi}_l(w_i, w_j)|^2 F_0(dw_i) F_0(dw_j) \rightarrow_p 0,$$

$$\text{or ii) } \sqrt{n} \binom{n}{2}^{-1} \sum_{(i,j) \in I_l} |\hat{\xi}_l(W_i, W_j)| \rightarrow_p 0, \text{ or iii) } \sqrt{n} \binom{n}{2}^{-1} \sum_{(i,j) \in I_l} \hat{\xi}_l(W_i, W_j) \rightarrow_p 0.$$

These are rate conditions on the remainder term $\hat{\xi}_l(w_i, w_j)$. For ϕ in Equation (3.4) in the main text, the interaction term has the form

$$\hat{\xi}_l(w_i, w_j) = (\hat{\alpha}_l(x_i, x_j) - \alpha_0(x_i, x_j)) (c_1 \hat{\gamma}_l(X_i) + c_2 \hat{\gamma}_l(X_j) - c_1 \gamma_0(X_i) - c_2 \gamma_0(X_j)).$$

Therefore, Assumption 5 follows for first steps satisfying orthogonality restrictions if $\sqrt{n}\|\hat{\alpha}_l - \alpha_0\| \|\hat{\gamma}_l - \gamma_0\| = o_p(1)$. This is a product rate condition which allows for nuisance estimators to converge at slower rates as long as the product converges at \sqrt{n} -rate. Define now $\bar{\psi}(\gamma, \alpha, \theta) \equiv \mathbb{E}[\psi(W_i, W_j, \gamma, \alpha, \theta)]$. Henceforth, C is a generic positive constant that may change from expression to expression.

Assumption 6 For each $l = 1, \dots, L$ and θ , i) $\int \int \phi(w_i, w_j, \gamma_0, \hat{\alpha}_l, \theta) F_0(dw_i) F_0(dw_j) = 0$ with probability approaching one; and either ii) $\|\hat{\gamma}_l - \gamma_0\| = o_p(n^{-1/4})$ and $|\bar{\psi}(\gamma, \alpha_0, \theta_0)| \leq C\|\gamma - \gamma_0\|^2$ for all γ with $\|\gamma - \gamma_0\|$ small enough; or iii) $\sqrt{n}\bar{\psi}(\hat{\gamma}_l, \alpha_0, \theta_0) \rightarrow_p 0$.

Assumption 6 (i) incorporates the global robustness property of α and is in most cases easy to check by inspection of ϕ . For example, it holds for ϕ in Equation (3.4) in the main text. Assumption 6 (ii) and (iii) are small bias conditions. For a parameter of the form $\theta_0 = \mathbb{E}[\delta(X_i, X_j, \gamma_0)(c_1\gamma_0(X_i) + c_2\gamma_0(X_j))]$, Assumption 6 (ii) and (iii) hold if $\|\hat{\gamma}_l - \gamma_0\| = o_p(n^{-1/4})$ and $\|\delta(\cdot, \gamma) - \delta(\cdot, \gamma_0)\| \leq C\|\gamma - \gamma_0\|$.

Lemma 4 If Assumptions 4-6 are satisfied then equation (2.1) holds.

Lemma 4 is a key asymptotic result which implies asymptotic normality for the debiased estimator by standard U-statistics arguments when the identifying moment condition is non-degenerate.

For valid inference we also need convergence of the asymptotic variance estimators. To simplify the computation we implemented the variance estimator without cross-fitting, so standard U-statistics formulas are valid. Consistency of the variance follows under standard L_2 consistency of first steps. Define $\hat{g}_{ij} = g(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \hat{\theta})$ and $g_{ij} = g(W_i, W_j, \gamma_0, \alpha_0, \theta_0)$. Then define the following leave-one out average $\hat{g}_{-i} = (n-1)^{-1} \sum_{j \neq i} \hat{g}_{ij}$ and $g_{-i} = (n-1)^{-1} \sum_{j \neq i} g_{ij}$, let $\hat{\phi}_{-i}$ and ϕ_{-i} be defined in the same way.

Lemma 5 If $\mathbb{E}[|\psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)|^2] < \infty$, $n^{-1} \sum_{i=1}^n |\hat{g}_{-i} - g_{-i}|^2 \rightarrow_p 0$ and $n^{-1} \sum_{i=1}^n |\hat{\phi}_{-i} - \phi_{-i}|^2 \rightarrow_p 0$, then $\hat{\Sigma} \rightarrow_p \Sigma$.

We also need convergence of the Jacobian of the moment condition $\hat{B} \rightarrow_p B$. Define $\tilde{\psi}_{ij} = \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \theta_0)$.

Assumption 7 B exists and there is a neighborhood \mathcal{N} of θ_0 such that i) $\|\hat{\gamma} - \gamma_0\| \rightarrow_p 0$; ii) for all $\|\gamma - \gamma_0\|$ and $\|\alpha - \alpha_0\|$ small enough $\psi(W_i, W_j, \gamma, \alpha, \theta)$ is differentiable in θ on \mathcal{N} with probability approaching one and there is $C > 0$ and $d(W_i, W_j, \gamma, \alpha)$ such that $\mathbb{E}[d(W_i, W_j, \gamma, \alpha)] \leq C$ and such that for $\theta \in \mathcal{N}$ and $\|\gamma - \gamma_0\|$ and $\|\alpha - \alpha_0\|$ small enough

$$\left| \frac{\partial \psi(W_i, W_j, \gamma, \alpha, \theta)}{\partial \theta} - \frac{\partial \psi(W_i, W_j, \gamma, \alpha, \theta_0)}{\partial \theta} \right| \leq d(W_i, W_j, \gamma, \alpha) |\theta - \theta_0|^{1/C}$$

iii) For each k , $\mathbb{E}[|\partial\tilde{\psi}_{ij}/\partial\theta_k - \partial\psi_{ij}/\partial\theta_k|] \rightarrow_p 0$.

Lemma 6 *If Assumption 7 is satisfied and if $\bar{\theta} \rightarrow_p \theta_0$ then $\partial\hat{\psi}(\bar{\theta})/\partial\theta \rightarrow_p B$.*

Now we are ready to give the main asymptotic result

Theorem 1 *If Assumptions 4-7 and conditions in Lemma 5 are satisfied, $\hat{\theta} \rightarrow_p \theta_0$ and $V = B^{-1}\Sigma B'^{-1}$ is nonsingular, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d \mathcal{N}(0, V).$$

Also, $\hat{V} \rightarrow_p V$.

Theorem 1 requires consistency of $\hat{\theta}$. In Appendix 2.2 we give conditions similar to the ones in this section under which $\hat{\theta}$ is consistent.

EXAMPLE 1 (Variance of fitted values, cont.): Assume for simplicity that Y_i and the fitted values $\hat{\gamma}_l(X_i)$ are bounded with probability one. Then, Assumptions 4-6 are easily verified and hold under the condition $\|\hat{\gamma}_l - \gamma_0\| = o_p(n^{-1/4})$. Furthermore, Assumption 7 also holds (trivially, as $\partial\psi/\partial\theta \equiv 1$). ■

REMARK 2 (On degeneracy of orthogonal moments). When the orthogonal (sample) moment is a degenerate U-statistic, $\Sigma = 0$ and hence $V = 0$. Therefore, our asymptotic distribution theory is useful only for non-degenerate orthogonal moments. We focus on this case because it is the most common one in applications, as for example in IOp. The analysis of the degenerate case is beyond the scope of this paper. ■

2.1 Asymptotic theory proofs

Proof of Lemma 4: Define

$$\begin{aligned} \hat{R}_{1,ij,l} &= g(W_i, W_j, \hat{\gamma}_l, \theta_0) - g(W_i, W_j, \gamma_0, \theta_0), & \hat{R}_{2,ij,l} &= \phi(W_i, W_j, \hat{\gamma}_l, \alpha_0, \theta_0) - \phi(W_i, W_j, \gamma_0, \alpha_0, \theta_0), \\ \hat{R}_{3,ij,l} &= \phi(W_i, W_j, \gamma_0, \hat{\alpha}_l, \theta_0) - \phi(W_i, W_j, \gamma_0, \alpha_0, \theta_0), & (i, j) &\in I_l. \end{aligned}$$

Then

$$g(W_i, W_j, \hat{\gamma}_l, \theta_0) + \phi(W_i, W_j, \hat{\gamma}_l, \hat{\alpha}_l, \theta_0) - \psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0) = \hat{R}_{1,ij,l} + \hat{R}_{2,ij,l} + \hat{R}_{3,ij,l} + \hat{\xi}_l(W_i, W_j). \quad (2.2)$$

Let N_l^c be the observations not in I_l and note that $\hat{\gamma}_l$ and $\hat{\alpha}_l$ depend only on N_l^c . Hence,

$$\mathbb{E}[\hat{R}_{1,ij,l}|N_l^c] = \int \int g(w_i, w_j, \hat{\gamma}_l, \theta_0) F_0(dw_i) F_0(dw_j), \quad (2.3)$$

$$\mathbb{E}[\hat{R}_{2,ij,l}|N_l^c] = \int \int \phi(w_i, w_j, \hat{\gamma}_l, \alpha_0, \theta_0) F_0(dw_i) F_0(dw_j), \quad (2.4)$$

$$\mathbb{E}[\hat{R}_{3,ij,l}|N_l^c] = \int \int \phi(w_i, w_j, \gamma_0, \hat{\alpha}_l, \theta_0) F_0(dw_i) F_0(dw_j) = 0, \quad (2.5)$$

where the two first equalities follow since $\mathbb{E}[g(W_i, W_j, \gamma_0, \theta_0)] = 0$ and $\mathbb{E}[\phi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)] = 0$ and in the third equality we use Assumption 6(i). Since pairs in I_l are dependent only when one or two of the members of the pair coincide (also we omit the fact that we are dealing with vectors since the convergence of the vector is the convergence of its elements)

$$\begin{aligned} & \mathbb{E} \left[\left(\sqrt{n} \binom{n}{2}^{-1} \sum_{(i,j) \in I_l} (\hat{R}_{1,ij,l} - \mathbb{E}(\hat{R}_{1,ij,l}|N_l^c)) \right)^2 \middle| N_l^c \right] = \\ & n \binom{n}{2}^{-2} \left[\kappa_{2,l} \text{Var}(\hat{R}_{1,ij,l}|N_l^c) + \kappa_{1,l} \text{Cov}(\hat{R}_{1,ij,l}, \hat{R}_{1,ik,l}|I_l^c) \right], \end{aligned}$$

where $\kappa_{r,l}$ is the number of ways of choosing a pair of pairs in I_l that have $r = 1, 2$, elements in common. These depend on the way you partition the data but generally $n \binom{n}{2}^{-2} \kappa_{2,l} \rightarrow 0$ as $n \rightarrow \infty$ and $n \binom{n}{2}^{-2} \kappa_{1,l} \leq 4$. For instance, with our partition (see Supplementary Material), $\kappa_{1,l} = \binom{|C_k|}{2} 2(|C_k| - 2)$ and $\kappa_{2,l} = \binom{|C_k|}{2}$ for the elements of the partition $\{C_k \times C_k : i < j\}$ for $k = 1, \dots, K$. For the elements of the partition $\{C_k \times C_m : m > k\}$ we have $\kappa_{1,l} = |C_k| |C_m| (|C_k| + |C_m|)$ and $\kappa_{2,l} = |C_k| |C_m|$. Since $|C_k| \leq n$ for $k = 1, \dots, K$, it follows that $n \binom{n}{2}^{-2} \kappa_{2,l} \rightarrow 0$ as $n \rightarrow \infty$ and $n \binom{n}{2}^{-2} \kappa_{1,l} \leq 4$. Hence

$$\begin{aligned} & n \binom{n}{2}^{-2} \left[\kappa_{2,l} \text{Var}(\hat{R}_{1,ij,l}|N_l^c) + \kappa_{1,l} \text{Cov}(\hat{R}_{1,ij,l}, \hat{R}_{1,ik,l}|I_l^c) \right] \\ & \leq 4 \text{Cov}(\hat{R}_{1,ij,l}, \hat{R}_{1,ik,l}|N_l^c) + o_P(1) \\ & \leq 4 \sqrt{\text{Var}(\hat{R}_{1,ij,l}|N_l^c) \text{Var}(\hat{R}_{1,ik,l}|N_l^c)} + o_P(1) \\ & \leq 4 \sqrt{\mathbb{E}(\hat{R}_{1,ij,l}^2|N_l^c) \mathbb{E}(\hat{R}_{1,ik,l}^2|N_l^c)} + o_P(1) \rightarrow_p 0, \end{aligned}$$

where the convergence in probability follows from Assumption 4(i). The same can be shown for $\hat{R}_{2,ij,l}$ with Assumption 4(ii) and for $\hat{R}_{3,ij,l}$ with Assumption 4(iii). By Assumption 6(i)

and the conditional Markov inequality

$$\sqrt{n} \binom{n}{2}^{-1} \sum_{(i,j) \in I_l} (\hat{R}_{1,ij,l} + \hat{R}_{2,ij,l} + \hat{R}_{3,ij,l} - \mathbb{E}[\hat{R}_{1,ij,l} + \hat{R}_{2,ij,l} | N_l^c]) \rightarrow_p 0.$$

By equations (2.3) and (2.4), $\mathbb{E}[\hat{R}_{1,ij,l} + \hat{R}_{2,ij,l} | N_l^c] = \mathbb{E}[\psi(W_i, W_j, \hat{\gamma}_l, \alpha_0, \theta_0)]$ (with $\hat{\gamma}_l$ fixed due to the conditioning on N_l^c). Therefore, by Assumption 6

$$\left| \sqrt{n} \binom{n}{2}^{-1} \sum_{(i,j) \in I_l} \mathbb{E}(\hat{R}_{1,ij,l} + \hat{R}_{2,ij,l} | N_l^c) \right| \leq 2\sqrt{n} |\bar{\psi}(\hat{\gamma}_l, \alpha_0, \theta_0)| \rightarrow_p 0.$$

By the triangle inequality

$$\sqrt{n} \binom{n}{2}^{-1} \sum_{(i,j) \in I_l} (\hat{R}_{1,ij,l} + \hat{R}_{2,ij,l} + \hat{R}_{3,ij,l}) \rightarrow_p 0.$$

Also, by Assumption 5

$$\sqrt{n} \binom{n}{2}^{-1} \sum_{(i,j) \in I_l} \hat{\xi}_l(W_i, W_j) \rightarrow_p 0.$$

Hence, by the triangle inequality and equation (2.2)

$$\begin{aligned} & \sqrt{n} \binom{n}{2}^{-1} \sum_{(i,j) \in I_l} \psi(W_i, W_j, \hat{\gamma}_l, \hat{\alpha}_l, \theta_0) - \psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0) \\ &= \sum_{l=1}^L \sqrt{n} (\hat{R}_{1,ij,l} + \hat{R}_{2,ij,l} + \hat{R}_{3,ij,l} + \hat{\xi}_l(W_i, W_j)) \rightarrow_p 0. \end{aligned}$$

■

Proof of Lemma 5: Let $\hat{\psi}_{ij} = \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \hat{\theta})$, $\psi_{ij} = \psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)$ and define $\hat{\chi}_i = (n-1)^{-1} \sum_{j \neq i} \hat{\psi}_{ij}$, $\tilde{\chi}_i = (n-1)^{-1} \sum_{j \neq i} \psi_{ij}$ and $\chi_i = \mathbb{E}[\psi_{ij} | W_i]$. Without loss of generality we focus on the scalar case. Note that $\hat{\Sigma} = 4n^{-1} \sum_{i=1}^n \hat{\chi}_i^2$ and we want to show that $\hat{\Sigma} \rightarrow_p 4\mathbb{E}[\chi_i^2] = \Sigma$. To do this note that

$$\frac{1}{n} \sum_{i=1}^n (\hat{\chi}_i - \tilde{\chi}_i + \tilde{\chi}_i)^2 = \frac{1}{n} \sum_{i=1}^n \tilde{\chi}_i^2 + \frac{1}{n} \sum_{i=1}^n (\hat{\chi}_i - \tilde{\chi}_i)^2 + \frac{2}{n} \sum_{i=1}^n (\hat{\chi}_i - \tilde{\chi}_i) \tilde{\chi}_i.$$

The first term in the RHS goes in probability to $\mathbb{E}[\chi_i^2]$ by standard U-statistic theory. By

Cauchy-Schwartz the third term can be bounded as

$$\frac{2}{n} \sum_{i=1}^n (\hat{\chi}_i - \tilde{\chi}_i) \tilde{\chi}_i \leq 2 \left(\frac{1}{n} \sum_{i=1}^n (\hat{\chi}_i - \tilde{\chi}_i)^2 \right)^{\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^n \tilde{\chi}_i^2 \right)^{\frac{1}{2}}.$$

Hence, if we show that $n^{-1} \sum_{i=1}^n |\hat{\chi}_i - \tilde{\chi}_i|^2 \rightarrow_p 0$ we have that $n^{-1} \sum_{i=1}^n \hat{\chi}_i^2 = \mathbb{E}[\chi_i^2] + o_p(1)$. It follows by the C_r inequality and the assumptions in the lemma that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n |\hat{\chi}_i - \tilde{\chi}_i|^2 &= \frac{1}{n} \sum_{i=1}^n \left| \frac{1}{n-1} \sum_{j \neq i} [\hat{g}_{ij} - g_{ij} + \hat{\phi}_{ij} - \phi_{ij}] \right|^2 \\ &\leq \frac{2}{n} \sum_{i=1}^n |\hat{g}_{-i} - g_{-i}|^2 + \frac{2}{n} \sum_{i=1}^n |\hat{\phi}_{-i} - \phi_{-i}|^2 \rightarrow_p 0. \end{aligned}$$

■

Proof of Lemma 6: Define

$$\hat{B} = \binom{n}{2}^{-1} \sum_{i < j} \partial \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \bar{\theta}) / \partial \theta, \quad \tilde{B} = \binom{n}{2}^{-1} \sum_{i < j} \partial \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \theta_0) / \partial \theta.$$

By ii), with probability approaching 1

$$\mathbb{E} \left[\binom{n}{2}^{-1} \sum_{i < j} d(W_i, W_j, \hat{\gamma}, \hat{\alpha}) \right] = \mathbb{E}[d(W_i, W_j, \hat{\gamma}, \hat{\alpha})] \leq C,$$

By Markov inequality, $\binom{n}{2}^{-1} \sum_{i < j} d(W_i, W_j, \hat{\gamma}, \hat{\alpha}) = O_p(1)$. So by ii) and $\bar{\theta} \rightarrow_p \theta_0$

$$\begin{aligned} \left| \hat{B} - \tilde{B} \right| &\leq \binom{n}{2}^{-1} \sum_{i < j} \left| \frac{\partial \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \bar{\theta})}{\partial \theta} - \frac{\partial \psi(W_i, W_j, \hat{\gamma}, \hat{\alpha}, \theta_0)}{\partial \theta} \right| \\ &\leq \binom{n}{2}^{-1} \sum_{i < j} d(W_i, W_j, \hat{\gamma}, \hat{\alpha}) |\bar{\theta} - \theta_0|^{1/C} \\ &= O_p(1) o_p(1) \rightarrow_p 0. \end{aligned}$$

It follows from Assumption 7 (iii) and Markov Inequality that

$$\left| \tilde{B} - \binom{n}{2}^{-1} \sum_{i < j} \partial \psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0) / \partial \theta \right| \rightarrow_p 0$$

and $\binom{n}{2}^{-1} \sum_{i < j} \partial \psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0) / \partial \theta \rightarrow_p B$ by usual U-statistics theory. Hence, the result follows from the triangle inequality. ■

Proof of Theorem 1: Let $\hat{\theta}(\gamma_0, \alpha_0)$ be the solution to the cross-fitted orthogonal sample moment if (γ_0, α_0) are known. It follows from Lemma 4 that $\hat{\theta}$ and $\hat{\theta}(\gamma_0, \alpha_0)$ are asymptotically equivalent. By the mean value theorem

$$\begin{aligned} 0 &= \sqrt{n} \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \psi(W_i, W_j, \gamma_0, \alpha_0, \hat{\theta}(\gamma_0, \alpha_0)) \\ &= \sqrt{n} \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0) \\ &\quad + \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \psi_\theta(W_i, W_j, \gamma_0, \alpha_0, \bar{\theta}) \sqrt{n} (\hat{\theta}(\gamma_0, \alpha_0) - \theta_0), \end{aligned}$$

where ψ_θ is the derivative with respect to θ and $\bar{\theta}$ is some intermediate value. Let $Q_n(\theta) \equiv \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \psi(W_i, W_j, \gamma_0, \alpha_0, \theta)$ and let $\binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \psi_\theta(W_i, W_j, \gamma_0, \alpha_0, \bar{\theta}) = M_n$. Then,

$$\begin{aligned} \sqrt{n}(\hat{\theta}(\gamma_0, \alpha_0) - \theta_0) &= M_n^{-1} \sqrt{n} Q_n(\theta_0) = - \left(M_n^{-1} - B^{-1} + B^{-1} \right) \sqrt{n} Q_n(\theta_0) \\ &= -B^{-1} \sqrt{n} Q_n(\theta_0) - \left(M_n^{-1} - B^{-1} \right) \sqrt{n} Q_n(\theta_0) \\ &= -B^{-1} \sqrt{n} Q_n(\theta_0) + o_p(1) O_p(1) \rightarrow_d \mathcal{N}(0, V), \end{aligned}$$

where

$$V = B^{-1} \text{Var} \left(\mathbb{E}[\psi(W_i, W_j, \gamma_0, \alpha_0, \theta_0) | W_i] \right) B'^{-1}.$$

The normality result follows from standard U-statistic theory (see Theorem 12.3 in [Van der Vaart \(2000\)](#)) and Slutsky's lemma. Hence $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d \mathcal{N}(0, V)$. Consistency of the variance estimator follows from Lemmas 5 and 6 and standard arguments.

■

2.2 Consistency

Re-define the interaction term in Section 2 as

$$\hat{\xi}_l(w_i, w_j, \theta) = \phi(w_i, w_j, \hat{\gamma}_l, \hat{\alpha}_l, \theta) - \phi(w_i, w_j, \gamma_0, \hat{\alpha}_l, \theta) - \phi(w_i, w_j, \hat{\gamma}_l, \alpha_0, \theta) + \phi(w_i, w_j, \gamma_0, \alpha_0, \theta).$$

Theorem 2 *If (i) $\mathbb{E}[g(W_i, W_j, \gamma_0, \theta)] = 0$ iff $\theta = \theta_0$, (ii) Θ is compact, (iii) $\int \int |g(w_i, w_j, \hat{\gamma}_l, \theta) - g(w_i, w_j, \gamma_0, \theta)| F_0(dw_i) F_0(dw_j) \rightarrow_p 0$ and $\mathbb{E}[|g(W_i, W_j, \gamma_0, \theta)|] < \infty$ for all $\theta \in \Theta$, (iv) there is*

$C > 0$ and $d(W_i, W_j, \gamma)$ such that for $\|\gamma - \gamma_0\|$ small enough and all $\tilde{\theta}, \theta \in \Theta$

$$|g(W_i, W_j, \gamma, \tilde{\theta}) - g(W_i, W_j, \gamma, \theta)| \leq d(W_i, W_j, \gamma) |\tilde{\theta} - \theta|^{1/C},$$

and $\mathbb{E}[d(W_i, W_j, \gamma)] < C$ and (v) Assumption 4 (ii), (iii), $\int \int |\hat{\xi}_l(w_i, w_j, \theta)| F_0(dw_i) F_0(dw_j) \rightarrow_p 0$ and $\mathbb{E}[\phi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)] < \infty$. Then $\hat{\theta} \rightarrow_p \theta_0$.

Proof: Define $\hat{g}(\theta) \equiv \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} g(W_i, W_j, \hat{\gamma}_l, \theta)$ and $\bar{g}(\theta) \equiv \mathbb{E}[g(W_i, W_j, \gamma_0, \theta)]$. It follows from the conditional Markov inequality and (iii) that $\hat{g}(\theta) \rightarrow_p \bar{g}(\theta)$ for all $\theta \in \Theta$. Let $\tilde{\phi}_{ij} \equiv \binom{n}{2}^{-1} \sum_{i < j} \phi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)$ and $\hat{\phi}_{ij}(\theta) \equiv \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \phi(W_i, W_j, \hat{\gamma}_l, \hat{\alpha}_l, \theta)$. In the notation of Lemma 4, $\hat{\phi}_{ij}(\theta) - \tilde{\phi}_{ij} = \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \hat{R}_{2,ij,l} + \hat{R}_{3,ij,l} + \hat{\xi}_l(W_i, W_j, \theta)$, so $\hat{\phi}_{ij}(\theta) - \tilde{\phi}_{ij} \rightarrow_p 0$ for all $\theta \in \Theta$ by Assumption (v) and the conditional Markov inequality. By consistency of U-statistics we have that $\tilde{\phi}_{ij} \rightarrow_p \mathbb{E}[\phi(W_i, W_j, \gamma_0, \alpha_0, \theta_0)] = 0$ so by the triangle inequality we have that $\hat{\phi}_{ij}(\theta) \rightarrow_p 0$ for all $\theta \in \Theta$. Therefore, defining $\hat{\psi}_{ij,l}(\theta) \equiv \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \psi(W_i, W_j, \hat{\gamma}_l, \hat{\alpha}_l, \theta)$, we have that $\hat{\psi}_{ij,l}(\theta) = \hat{g}(\theta) + o_p(1)$. By triangle inequality and (iv) we know that with probability approaching one

$$|\hat{g}(\hat{\theta}) - \hat{g}(\theta)| \leq \binom{n}{2}^{-1} \sum_{i < j} |g(W_i, W_j, \hat{\gamma}_l, \hat{\theta}) - g(W_i, W_j, \hat{\gamma}_l, \theta)| \leq \underbrace{\binom{n}{2}^{-1} \sum_{i < j} d(W_i, W_j, \hat{\gamma}_l)}_{\equiv \hat{M}_l} |\hat{\theta} - \theta|^{1/C},$$

and by (iv) and the conditional Markov inequality $\hat{M}_l = O_p(1)$. By Corollary 2.2 in Newey (1991) we have that $\sup_{\theta \in \Theta} |\hat{\psi}(\theta) - \bar{g}(\theta)| = o_p(1)$. We also know that $\bar{g}(\theta)$ is continuous by (iv). So the conclusion follows from the proof of Theorem 2.6 in Newey and McFadden (1994) applied to the Hájek projection of $\hat{g}(\hat{\theta})$. ■

3 Further Applications

3.1 Bipartite ranking problem

Consider data $W_i = (Y_i, X_i) \in \{-1, +1\} \times \mathbb{R}^k$. We want a rule $r(X_i, X_j) \in \{-1, +1\}$ which equals 1 if $Y_i = 1$ with larger probability than $Y_j = 1$, and equals -1 otherwise. The probability of committing a mistake (Ranking Risk) is

$$L(r) = P[r(X_i, X_j)(Y_i - Y_j) < 0].$$

It can be shown that the optimal rule r^* , i.e. $L(r^*) \leq L(r)$ for any r , is

$$r^* = 2\mathbb{I}(\gamma_0(X_i) > \gamma_0(X_j)) - 1,$$

where $\gamma_0(x) = P(Y_i = 1|X_i = x)$. Our parameter of interest is the Optimal Risk $L(r^*)$ which can be shown to be (see [Cl emen on and Robbiano \(2011\)](#))

$$\theta_0 = L(r^*) = \frac{1}{2} \mathbb{E} \left[(Y_i - Y_j)^2 - |\gamma_0(X_i) - \gamma_0(X_j)| \right].$$

Minimization of the Ranking Risk can be shown to be equivalent to maximization of the AUC (Area under the Receiver Operating Characteristic Curve ROC) which is ubiquitous in classification and biomedical problems. The identifying moment function is

$$g(w_i, w_j, \gamma, \theta) = \frac{1}{2} \left[(y_i - y_j)^2 - |\gamma(x_i) - \gamma(x_j)| \right] - \theta.$$

From Lemma 1, a U-FSIF for this identifying quadratic moment function is

$$\phi(w_i, w_j, \gamma_0) = -\frac{1}{2} \alpha_0(x_i, x_j) (y_i - y_j - \gamma_0(x_i) + \gamma_0(x_j)),$$

where $\alpha_0(x_i, x_j)$ is the same as for IOp. Adding the U-FSIF to the original identifying moment gives the debiased estimator

$$\hat{\theta} = \frac{1}{n(n-1)} \sum_{l=1}^L \sum_{(i,j) \in I_l} (Y_i - Y_j)^2 - |\hat{\gamma}_l(X_i) - \hat{\gamma}_l(X_j)| + \hat{\alpha}_l(X_i, X_j) (Y_i - Y_j - \hat{\gamma}_l(X_i) + \hat{\gamma}_l(X_j)).$$

If $\Gamma = L_2$ and $\mathcal{S} = L_2$, the estimator simplifies to

$$\hat{\theta} = \frac{1}{n(n-1)} \sum_{l=1}^L \sum_{(i,j) \in I_l} (Y_i - Y_j) (Y_i - Y_j - \delta(X_i, X_j, \hat{\gamma}_l)),$$

and the asymptotic variance can be estimated as

$$\hat{V} = \frac{2}{n(n-1)^2} \sum_{i=1}^n \left[\sum_{j \neq i} \left((Y_i - Y_j) (Y_i - Y_j - \delta(X_i, X_j, \hat{\gamma}_{kk'})) - \hat{\theta} \right)^2 \right].$$

The asymptotic theory for this example follows from the example of Inequality of Opportunity.

3.2 Treatment Effects

Let $W_i = (Y_i, D_i, X_i)$ and $Z_i = (Y_i(1), Y_i(0))$ where $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. Mao (2018) defines the following contrast parameter

$$\theta_0^h = \int \int h(y_i, y_j) F_{(1)}(dy_i) F_{(0)}(dy_j).$$

If $h(y_i, y_j) = y_i - y_j$ this coincides with the usual average treatment effect $\mathbb{E}[Y_i(1) - Y_i(0)]$. However, for nonlinear h it is not true in general that $\theta_0^h = \mathbb{E}[h(Y_i(1), Y_i(0))]$. Also, by picking a nonlinear *contrast function* h we can estimate different (counterfactual) treatment effects. Let $\gamma_0(X_i)$ be the population fitted values of D_i given X_i , i.e the propensity score. Throughout, we assume unconfoundedness and non-overlapping

$$(Y_i(1), Y_i(0)) \perp D_i | X_i \text{ and } \epsilon \leq \gamma_0(X_i) \leq 1 - \epsilon \text{ a.s., } \epsilon > 0. \quad (3.1)$$

Letting

$$g(w_i, w_j, \gamma, \theta) = \frac{1}{2} \left[\frac{d_i(1 - d_j)h(y_i, y_j)}{\gamma(x_i)(1 - \gamma(x_j))} + \frac{d_j(1 - d_i)h(y_j, y_i)}{\gamma(x_j)(1 - \gamma(x_i))} \right] - \theta,$$

one can show that under (3.1)

$$\mathbb{E}[g(W_i, W_j, \gamma_0, \theta)] = 0 \text{ iff } \theta = \theta_0^h.$$

Define now

$$\delta(w_i, w_j, \gamma) = \frac{1}{2} \left[\frac{d_j(1 - d_i)h(y_j, y_i)}{\gamma(x_j)(1 - \gamma(x_i))^2} - \frac{d_i(1 - d_j)h(y_i, y_j)}{\gamma(x_i)^2(1 - \gamma(x_j))} \right].$$

Then, it can be shown that

$$\begin{aligned} \frac{d}{d\tau} \mathbb{E}[g(W_i, W_j, \gamma(F_\tau), \theta)] &= \frac{d}{d\tau} \mathbb{E}[\delta(W_i, W_j, \gamma_0) \gamma_\tau(X_i)] + \frac{d}{d\tau} \mathbb{E}[\delta(W_j, W_i, \gamma_0) \gamma_\tau(X_j)] \\ &= \frac{d}{d\tau} \mathbb{E}[\alpha_0(X_i) \gamma_\tau(X_i)] + \frac{d}{d\tau} \mathbb{E}[\alpha_0(X_j) \gamma_\tau(X_j)], \end{aligned}$$

where $\alpha_0(X_i) = \Pi_\Gamma \mathbb{E}[\delta_{ij}(\gamma_0) | X_i]$. Then by the same arguments as in the previous examples

$$\phi(w_i, w_j, \gamma, \alpha) = \alpha(x_i)(d_i - \gamma(x_i)) + \alpha(x_j)(d_j - \gamma(x_j)).$$

The debiased estimator is

$$\hat{\theta}^h = \binom{n}{2}^{-1} \sum_{l=1}^L \sum_{(i,j) \in I_l} \frac{1}{2} \left[\frac{D_i(1 - D_j)h(Y_i, Y_j)}{\hat{\gamma}_l(X_i)(1 - \hat{\gamma}_l(X_j))} + \frac{D_j(1 - D_i)h(Y_j, Y_i)}{\hat{\gamma}_l(X_j)(1 - \hat{\gamma}_l(X_i))} + 2\phi(W_i, W_j, \hat{\gamma}_l, \hat{\alpha}_l) \right],$$

where $\hat{\alpha}_l$ is an estimator for α_0 . In the joint nonparametric case, we can set $\alpha_0 = \delta(\cdot, \gamma_0)$ and $\hat{\alpha}_l(\cdot) = \delta(\cdot, \hat{\gamma}_l)$, which leads to simpler implementations (as there is no need to estimate the conditional distribution of $h(Y_i, Y_j)$ given X_i). The variance is estimated as usual

$$\hat{V} = \frac{2}{n(n-1)^2} \sum_{i=1}^n \left[\sum_{j \neq i} \left(g(W_i, W_j, \hat{\gamma}) + \phi(W_i, W_j, \hat{\gamma}, \hat{\alpha}) \right)^2 \right].$$

4 Simulations

In the tables below we see the results of the simulations for $\sigma \in \{0.2, 0.3\}$ for such lower StN ratios noisier estimates such as Ridge start malfunctioning and well-performing estimators such as lasso or CIF need more observations before getting good results.

$\sigma = 0.2$		Gini of the FVs							
		Plug in Lasso		Debiased Lasso		Plug in Ridge		Debiased Ridge	
		Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage
n = 100		-0.012	0.611	-0.050	0.409	-0.080	0.000	-0.031	0.605
n = 500		-0.016	0.292	-0.009	0.730	-0.074	0.000	-0.010	0.687
n = 1000		-0.016	0.083	-0.003	0.909	-0.020	0.042	-0.009	0.611
n = 3000		-0.009	0.060	-0.001	0.935	-0.005	0.427	-0.005	0.554
		Plug in RF		Debiased RF		Plug in CIF		Debiased CIF	
		Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage
n = 100		-0.026	0.280	-0.023	0.730	-0.070	0.000	-0.023	0.692
n = 500		-0.020	0.060	-0.005	0.865	-0.033	0.000	-0.004	0.901
n = 1000		-0.020	0.002	-0.003	0.897	-0.026	0.000	-0.002	0.925
n = 3000		-0.022	0.000	-0.001	0.913	-0.022	0.000	-0.001	0.923
		Plug in XGBoost		Debiased XGBoost		Plug in Catboost		Debiased Catboost	
		Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage
n = 100		0.030	0.427	-0.060	0.252	-0.014	0.520	-0.048	0.337
n = 500		0.015	0.292	-0.019	0.335	0.002	0.685	-0.007	0.794
n = 1000		0.009	0.361	-0.009	0.593	0.002	0.700	-0.003	0.897
n = 3000		0.005	0.367	-0.002	0.907	0.002	0.687	-0.001	0.946

Table 4: Simulation based on 500 Monte Carlo iterations, true value for the Gini of the fitted values is 0.18.

$\sigma = 0.3$		Gini of the FVs							
	Plug in Lasso		Debiased Lasso		Plug in Ridge		Debiased Ridge		
	Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage	
n = 100	-0.041	0.395	-0.085	0.264	-0.073	0.000	-0.044	0.621	
n = 500	-0.028	0.115	-0.019	0.508	-0.066	0.000	-0.017	0.563	
n = 1000	-0.027	0.022	-0.009	0.722	-0.035	0.002	-0.015	0.427	
n = 3000	-0.017	0.006	-0.002	0.913	-0.025	0.000	-0.009	0.355	
	Plug in RF		Debiased RF		Plug in CIF		Debiased CIF		
	Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage	
n = 100	-0.017	0.552	-0.035	0.716	-0.067	0.002	-0.034	0.659	
n = 500	-0.016	0.212	-0.011	0.788	-0.032	0.004	-0.008	0.839	
n = 1000	-0.017	0.034	-0.007	0.806	-0.024	0.000	-0.004	0.879	
n = 3000	-0.020	0.000	-0.002	0.889	-0.021	0.000	-0.002	0.917	
	Plug in XGBoost		Debiased XGBoost		Plug in Catboost		Debiased Catboost		
	Bias	Coverage	Bias	Coverage	Bias	Coverage	Bias	Coverage	
n = 100	0.059	0.042	-0.082	0.298	-0.015	0.466	-0.068	0.317	
n = 500	0.028	0.022	-0.036	0.095	0.002	0.563	-0.015	0.657	
n = 1000	0.017	0.067	-0.020	0.206	0.002	0.560	-0.007	0.794	
n = 3000	0.008	0.127	-0.006	0.673	0.002	0.565	-0.001	0.946	

Table 5: Simulation based on 500 Monte Carlo iterations, true value for the Gini of the fitted values is 0.18.

5 Cross-Fitting

We create a partition $\mathcal{C} = \{C_1, \dots, C_K\}$ of the set $\mathcal{N} = \{1, \dots, n\}$. Then we create partition $\mathcal{I} = \{I_1, \dots, I_L\}$ of the set $\{(i, j) \in \mathcal{N}^2 : i < j\}$ as depicted in Figure 7 for $n = 21$ and $K = 3$. This partition results in $L = K(K + 1)/2$ blocks. For example, in Figure 7 when pairs in I_1 are used to compute the sample orthogonal moment, the nuisance parameters are estimated with observations in (C_2, C_3) .

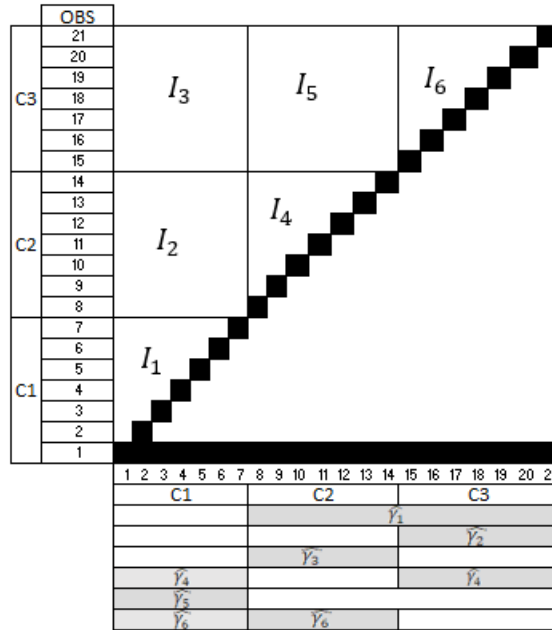


Figure 7: $n = 21$, $L = 3$

This is nothing else than a partition of \mathcal{C}^2 in squares intersected with the set $\{i < j : i, j \in \mathcal{N}\}$.

References

- AIYAR, S. AND C. EBEKE (2020): “Inequality of opportunity, inequality of income and economic growth,” *World Development*, 136, 105115.
- ALESINA, A. AND E. LA FERRARA (2005): “Preferences for redistribution in the land of opportunities,” *Journal of public Economics*, 89, 897–931.
- ARCONES, M. A. AND E. GINÉ (1993): “Limit theorems for U-processes,” *The Annals of Probability*, 1494–1542.
- ATHEY, S., J. TIBSHIRANI, AND S. WAGER (2019): “Generalized random forests,” *The Annals of Statistics*, 47, 1148–1178.
- BELLONI, A. AND V. CHERNOZHUKOV (2011): “l1-penalized quantile regression in high-dimensional sparse models,” .
- (2013): “Least squares after model selection in high-dimensional sparse models,” .
- BICKEL, P. J. (1982): “On adaptive estimation,” *The Annals of Statistics*, 647–671.
- BLUNDELL, R. W. AND J. L. POWELL (2004): “Endogeneity in semiparametric binary response models,” *The Review of Economic Studies*, 71, 655–679.
- BRUNORI, P., P. HUFÉ, AND D. MAHLER (2021): “The roots of inequality: Estimating inequality of opportunity from regression trees and forests,” *IZA Discussion Paper No. 14689*.
- BRUNORI, P. AND G. NEIDHÖFER (2021): “The evolution of inequality of opportunity in Germany: A machine learning approach,” *Review of Income and Wealth*, 67, 900–927.
- BRUNORI, P., F. PALMISANO, AND V. PERAGINE (2019a): “Inequality of opportunity in sub-Saharan Africa,” *Applied Economics*, 51, 6428–6458.
- BRUNORI, P., V. PERAGINE, AND L. SERLENGA (2019b): “Upward and downward bias when measuring inequality of opportunity,” *Social Choice and Welfare*, 52, 635–661.
- CARRANZA, R. (2020): “Inequality of outcomes, inequality of opportunity, and economic growth,” Tech. rep., ECINEQ, Society for the Study of Economic Inequality.
- (2022): “Upper and Lower Bound Estimates of Inequality of Opportunity: A Cross-National Comparison for Europe,” *Review of Income and Wealth*.

- CHANG, Y., S. N. DURLAUF, S. LEE, AND J. Y. PARK (2023): “A Trajectories-Based Approach to Measuring Intergenerational Mobility,” Working Paper 31020, National Bureau of Economic Research.
- CHEN, T. AND C. GUESTRIN (2016): “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
- CHEN, X. (2007): “Large sample sieve estimation of semi-nonparametric models,” *Handbook of econometrics*, 6, 5549–5632.
- CHEN, X. AND H. WHITE (1999): “Improved rates and asymptotic normality for non-parametric neural network estimators,” *IEEE Transactions on Information Theory*, 45, 682–691.
- CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal*, 21, C1–C68.
- CHERNOZHUKOV, V., J. C. ESCANCIANO, H. ICHIMURA, W. K. NEWEY, AND J. M. ROBINS (2022): “Locally robust semiparametric estimation,” *Forthcoming Econometrica*.
- CHIANG, H. D., K. KATO, Y. MA, AND Y. SASAKI (2021a): “Multiway cluster robust double/debiased machine learning,” *Journal of Business & Economic Statistics*, 1–11.
- CHIANG, H. D., Y. MA, J. RODRIGUE, AND Y. SASAKI (2021b): “Dyadic Double/Debiased Machine Learning for Analyzing Determinants of Free Trade Agreements,” *arXiv preprint arXiv:2110.04365*.
- CLÉMENÇON, S., G. LUGOSI, AND N. VAYATIS (2008): “Ranking and empirical minimization of U-statistics,” *The Annals of Statistics*, 36, 844–874.
- CLÉMENÇON, S. AND S. ROBBIANO (2011): “Minimax learning rates for bipartite ranking and plug-in rules,” in *International Conference on Machine Learning*, 441–448.
- DOMÍNGUEZ, M. A. AND I. N. LOBATO (2004): “Consistent estimation of models defined by conditional moment restrictions,” *Econometrica*, 72, 1601–1615.
- DUCLOS, J.-Y., J. ESTEBAN, AND D. RAY (2004): “Polarization: concepts, measurement, estimation,” *Econometrica*, 72, 1737–1772.
- FARRELL, M. H., T. LIANG, AND S. MISRA (2021): “Deep neural networks for estimation and inference,” *Econometrica*, 89, 181–213.

- FERREIRA, F. H., C. LAKNER, M. A. LUGO, AND B. ÖZLER (2018): “Inequality of opportunity and economic growth: how much can cross-country regressions really tell us?” *Review of Income and Wealth*, 64, 800–827.
- FERREIRA, F. H. AND V. PERAGINE (2016): “Individual responsibility and equality of opportunity,” in *The Oxford handbook of well-being and public policy*.
- FLEURBAEY, M. (1995): “Equal opportunity or equal social outcome?” *Economics & Philosophy*, 11, 25–55.
- HAMPEL, F. R. (1974): “The influence curve and its role in robust estimation,” *Journal of the american statistical association*, 69, 383–393.
- HANSEN, B. E. (2008): “Uniform convergence rates for kernel estimation with dependent data,” *Econometric Theory*, 24, 726–748.
- HEUCHENNE, C. AND A. JACQUEMAIN (2022): “Inference for monotone single-index conditional means: A Lorenz regression approach,” *Computational Statistics & Data Analysis*, 167, 107347.
- HONORÉ, B. AND J. POWELL (2005): “Pairwise difference estimators for nonlinear models,” in *Identification and Inference for Econometric Models W. Andrews and J. H. Stock*, Cambridge University Press, Cambridge.
- HOTHORN, T., K. HORNIK, AND A. ZEILEIS (2006): “Unbiased recursive partitioning: A conditional inference framework,” *Journal of Computational and Graphical statistics*, 15, 651–674.
- HUFE, P., R. KANBUR, AND A. PEICHL (2022a): “Measuring Unfair Inequality: Reconciling Equality of Opportunity and Freedom from Poverty,” *The Review of Economic Studies*, 89, 3345–3380.
- HUFE, P., A. PEICHL, P. SCHÜLE, AND J. TODOROVIĆ (2022b): “Fairness in Europe: A Multidimensional Comparison,” in *CESifo Forum*, München: ifo Institut-Leibniz-Institut für Wirtschaftsforschung an der . . . , vol. 23, 45–51.
- HUFE, P., A. PEICHL, AND D. WEISHAAR (2022c): “Lower and upper bound estimates of inequality of opportunity for emerging economies,” *Social Choice and Welfare*, 58, 395–427.
- ICHIMURA, H. AND W. K. NEWEY (2022): “The influence function of semiparametric estimators,” *Quantitative Economics*, 13, 29–61.

- KLAASSEN, C. A. (1987): “Consistent estimation of the influence function of locally asymptotically linear estimators,” *The Annals of Statistics*, 15, 1548–1562.
- KUECK, J., Y. LUO, M. SPINDLER, AND Z. WANG (2023): “Estimation and inference of treatment effects with L2-boosting in high-dimensional settings,” *Journal of Econometrics*, 234, 714–731.
- LEE, A. J. (2019): *U-statistics: Theory and Practice*, Routledge.
- LOU, Z., X. ZHANG, AND W. B. WU (2023): “High-dimensional analysis of variance in multivariate linear regression,” *Biometrika*, 110, 777–797.
- LOUNICI, K. (2008): “Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators,” .
- MAMMEN, E. AND A. B. TSYBAKOV (1999): “Smooth discrimination analysis,” *The Annals of Statistics*, 27, 1808–1829.
- MAO, L. (2018): “On causal estimation using U-statistics,” *Biometrika*, 105, 215–220.
- MARRERO, G. A. AND J. G. RODRÍGUEZ (2012): “Inequality of opportunity in Europe,” *Review of Income and Wealth*, 58, 597–621.
- (2013): “Inequality of opportunity and growth,” *Journal of development Economics*, 104, 107–122.
- NEWBY, W. K. (1991): “Uniform convergence in probability and stochastic equicontinuity,” *Econometrica: Journal of the Econometric Society*, 1161–1167.
- (1994): “The asymptotic variance of semiparametric estimators,” *Econometrica: Journal of the Econometric Society*, 1349–1382.
- NEWBY, W. K. AND D. MCFADDEN (1994): “Large sample estimation and hypothesis testing,” *Handbook of econometrics*, 4, 2111–2245.
- NOLAN, D. AND D. POLLARD (1988): “Functional limit theorems for U -processes,” *The Annals of Probability*, 16, 1291–1298.
- POWELL, J. L. (1987): “Semiparametric estimation of bivariate latent variable models,” *SSRI Workshop Series*.
- PROKHORENKOVA, L., G. GUSEV, A. VOROBEV, A. V. DOROGUSH, AND A. GULIN (2018): “CatBoost: unbiased boosting with categorical features,” *Advances in neural information processing systems*, 31.

- RAJARSHI MUKHERJEE, WHITNEY K NEWEY, J. M. R. (2017): “Semiparametric Efficient Empirical Higher Order Influence Function Estimators,” .
- RAMOS, X. AND D. VAN DE GAER (2016): “Approaches to inequality of opportunity: Principles, measures and evidence,” *Journal of Economic Surveys*, 30, 855–883.
- ROBINS, J. M., L. LI, R. MUKHERJEE, E. T. TCHETGEN, AND A. VAN DER VAART (2017): “Minimax estimation of a functional on a structured high-dimensional model,” *The Annals of Statistics*, 45, 1951–1987.
- ROEMER, J. E. (1998): *Equality of opportunity*, Harvard University Press.
- ROEMER, J. E. AND A. TRANNOY (2016): “Equality of opportunity: Theory and measurement,” *Journal of Economic Literature*, 54, 1288–1332.
- SALAS-ROJO, P. AND J. G. RODRÍGUEZ (2022): “Inheritances and wealth inequality: a machine learning approach,” *The Journal of Economic Inequality*, 1–25.
- SASAKI, Y. AND T. URA (2021): “Estimation and inference for policy relevant treatment effects,” *Journal of Econometrics*.
- SCHICK, A. (1986): “On asymptotically efficient estimation in semiparametric models,” *The Annals of Statistics*, 1139–1151.
- SCHOLKOPF, B. AND A. J. SMOLA (2018): *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press.
- SHERMAN, R. P. (1994): “U-processes in the analysis of a generalized semiparametric regression estimator,” *Econometric theory*, 10, 372–395.
- TERSCHUUR, J. (2022): “Debiased Machine Learning Inequality in Europe,” *arXiv preprint arXiv:2212.02407*.
- TSYBAKOV, A. B. (2004): “Optimal aggregation of classifiers in statistical learning,” *The Annals of Statistics*, 32, 135–166.
- VAN DE GAER (1993): “Equality of opportunity and investment in human capital,” Phd thesis, Katholieke Universiteit Leuven.
- VAN DER LAAN, M. J. AND A. F. BIBAUT (2017): “Uniform consistency of the highly adaptive lasso estimator of infinite dimensional parameters,” *arXiv preprint arXiv:1709.06256*.
- VAN DER VAART, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

- VON MISES, R. V. (1947): “On the asymptotic distribution of differentiable statistical functions,” *The annals of mathematical statistics*, 18, 309–348.
- WAGER, S. AND G. WALTHER (2015): “Adaptive concentration of regression trees, with application to random forests,” *arXiv preprint arXiv:1503.06388*.
- WERNER, T. (2021): “A review on instance ranking problems in statistical learning,” *Machine Learning*, 1–49.
- WU, P., Y. HAN, T. CHEN, AND X. TU (2014): “Causal inference for Mann–Whitney–Wilcoxon rank sum and other nonparametric statistics,” *Statistics in medicine*, 33, 1261–1271.
- YAMAMURO, S. (1974): *Differential calculus in topological linear spaces*, vol. 374, Springer.
- YITZHAKI, S. AND I. OLKIN (1991): “Concentration indices and concentration curves,” *Lecture Notes-Monograph Series*, 19, 380–392.
- ZHANG, T. AND B. YU (2005): “Boosting with early stopping: Convergence and consistency,” .