Testing Welfare Improvability

2024-11-23

Abstract

Here is a placeholder for the abstract.

## Introduction

Potential Names - Comparing Algorithmic Impacts with Doubly-Robust Estimation - Statistical Inference for Comparing Treatment Assignment Algorithms - Making Informed Choices: A Framework for Evaluating the Impact of Competing Algorithms - Statistical Inference for Comparing Treatment Assignment Algorithms

## (old) Policy Evaluation

The use of algorithmic decision-making systems is becoming increasingly pervasive across many areas of society. There is a growing critical need for robust methods to evaluate their performance and impacts on those they make decisions about. From healthcare and criminal justice to financial services and education, these systems are increasingly shaping outcomes that profoundly affect individual lives and social structures. While there is great potential for algorithms to enhance efficiency and fairness, their use raises significant concerns about their impacts on well-being. However, the potential of algorithms to enhance efficiency and fairness should instil optimism about their future impact.

Algorithms have emerged as powerful and valuable tools for addressing complex economic decision problems, offering substantial benefits across various domains. Ludwig, Mullainathan, and Rambachan, in their study on “The Unreasonable Effectiveness of Algorithms,” argue that they can also provide a “free lunch in terms of public spending. For instance, in the criminal justice system, an algorithm applied to pretrial release decisions in New York City demonstrated the potential to reduce pretrial detentions by up to Without increasing failure rates (REF). In healthcare, an algorithmic approach to diagnosing heart attacks could potentially reduce unnecessary stress tests and catheterisations, leading to significant cost savings, potentially billions in Medicare costs annually.

Given the promise, we need robust methods to help us decide *which* algorithm to use.

The power of algorithms lies in their ability to extract signals from complex datasets, often outperforming human judgement in ranking and prediction tasks. This capability allows for more efficient allocation of resources and more accurate decision-making in various economic contexts. In education, Bergman et al. (2023) found that an algorithm for college course placement increased enrollments in college-level classes without compromising pass rates while also reducing disparities across racial and ethnic groups. In workplace safety regulation, Johnson et al. (2023) demonstrated that an algorithm could better predict which work sites will likely have future injuries, potentially preventing thousands of severe injuries and saving hundreds of millions of dollars in lost income.

However, it is crucial to note that the effectiveness of algorithms has its challenges. These promising results should not lead to immediate large-scale implementation but encourage further research and development in algorithmic solutions to policy problems. The need for this research is urgent and of utmost importance. Key challenges remain, such as understanding how decision-makers will respond to algorithmic tools in practice (Albright REF), addressing potential data drift over time, ensuring algorithms generalise across different contexts, and accurately assessing the impact of policies. This paper aims to contribute to addressing that last problem.

Recent years have witnessed substantial progress in developing frameworks for assessing the welfare implications of algorithmic decisions, mainly through the lens of treatment effects. This includes the empirical welfare maximisation literature, including the seminal work by and , which established rigorous foundations for policy learning and demonstrating the possibility of deriving optimal treatment assignment policies that maximise welfare under various constraints. The purpose of empirical welfare maximisation is to leverage treatment heterogeneity to maximise welfare using an algorithmic decision policy at an almost convergence rate. Athey and Wager’s work, rooted in the theory of semiparametric efficiency estimation , has opened new avenues for understanding and improving algorithmic decision-making processes. However, the existing literature on policy learning often focuses on asymptotic optimality within simplified policy classes, prioritizing theoretical guarantees over their practical applicability. While these contributions are invaluable, there remains a pressing need for more pragmatic approaches that can navigate the complexities of real-world algorithmic systems. This research aims to bridge the gap by proposing a novel framework for assessing welfare impacts of algorithmic treatment decisions and to help practitioners implement and find decision policies. There is concern around the finite-sample performance of some of this methods, with limited guarantees of optimally in finite samples.

Our approach introduces a testing procedure for welfare improvability that evaluates a status quo algorithm or policy against a class of proposed algorithmic policies. This method offers several key advantages over existing methods proposed in the empirical welfare maximization literature. First, it allows for greater flexibility in the policy class selection, with almost no constraints placed on the set of allowed policies a priori. This can accommodate large and complex state-of-the-art algorithms such as deep neural networks or large ensembles that may outperform simpler, analytically tractable policy sets in practice. Second, it yields interpretable results, in terms of treatment effects, that can be readily understood by policymakers, legal experts, and other non-technical stakeholders. Third, it has potential applications in legal and ethical contexts, providing a framework for demonstrating the absence of discriminatory practices or the impossibility of Pareto improvements in welfare for specific subgroups such as men and women.

Given sufficiently fast convergence of estimation of the nucciance parameters in the estimation of average treatment effects, we show that the proposed bootstrap procedure is consistent in its estimation of average welfare. This approach allows for the welfare ranking of policies (). The decision theoretic idea of optimising some population decision function using a sample analogue is known as the Empirical Risk Minimising principal in classification and statistical learning (Vapnik, 1998).

By adopting this methodology, we take a more nuanced view of algorithmic decision-making than traditional empirical welfare maximization approaches. Rather than focus solely on identifying the optimal policy within a restricted class, we provide a framework for evaluating and comparing complex algorithmic systems in terms of their welfare impacts. This shift in perspective opens up new possibilities for understanding and improving algorithmic decision-making in real-world contexts. As algorithmic systems become increasingly influential and critical in decision-making processes, the ability to rigorously assess their welfare impacts becomes paramount. This approach provides a practical tool for policymakers, legal professionals, researchers, and ethicists to evaluate the fairness and efficacy of algorithmic decisions. Moreover, it offers a means to detect and address potential biases or inefficiencies in existing systems, contributing to the development of more equitable and efficient algorithmic policies. UK Government’s Data Ethics Framework which asks practitioners to perform a self-assessment of their transparency, fairness, and accountability [18].

Furthermore, it has been argued that using complex models are often preferable to simple models [Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability]. The authors show that for every simple prediction function, there exists a more complex function that is strictly more equitable and more efficient. (Can the statement be made that for every simple policy class, there exists a more favorable complex policy class. What does learnability have to say about comparisons across policy classes?).

In the sections that follow, we first review the literature on policy evaluation, building on the seminal work on semi-parametric efficiency by . We then propose our testing procedure and show that we efficiently and unbiasedly estimate differences in welfare of different algorithmic decision-making policies and show the convergence of our bootstrap estimation procedure. Wed Monte Carlo studies results, showing how the method is used in practice, and finally, we apply this method to the real-world setting of [TO BE FILLED IN LATER].

By offering a pragmatic yet rigorous approach to assessing the welfare impacts of algorithmic decisions, this research aims to contribute to the ongoing dialogue about the role and use of algorithms. The tools presented here will help develop a more transparent, accountable, and welfare-enhancing use of algorithmic decision policies.

We assume that we have independent and identically distributed samples , where is an outcome we try to predict, is the observed treatment assignment, are a subject’s features, and is an (optional) instrument, and is interpreted as the utility resulting from the intervention. If is conditionally exogenous, then . It might also be the case that our target variable is equal to the measured utility .

A decision-making algorithm is is a mapping from a subject’s features to a decision .

We can define the causal effect of an intervention in terms of potential outcomes, where correspond to the utility of subjects under treatment .

When our treatment is binary, the utility under the algorithm is measured against no-treatment

In the policy learning literature, the policy is then usually evaluated as the regret of the algorithm relative to the best algorithm in class given by:

is usually constrained to be a limited class of functions, such as linear decision rules or shallow decision trees. This is done to contain the VC dimension of the policy class to make asymptotic regret guarantees tractable. In this research, no such restrictions are placed on the class of allowed potential decision algorithms.

As shown by Kitagawa and Tetenov (2018), if is exogenous with known treatment propensities and a suitable class of policies, inverse propensity weighting can be used to derive a policy whose regret decays with , with

Athey and Wager (2020) extend this to cases where treatment propensities are unknown and may need to be identified from operational data.

In the case of a continuous intervention, the utility of an infinitesimal intervention is given by

with regret defined similarly. An example of a continuous intervention would be price interventions.

In the case of using observational data, as opposed to using data generated from an experiment, we need to make assumptions about the data-generating process that allow for identification and estimation of expected utility .

The empirical welfare maximisation literature relies on controlling the size of the policy class , to make an estimation of a “best in class” policy realistic. This work takes a more pragmatic view of the problem in that we are more interested in testing if we are able to produce a decision-making algorithm that results in higher welfare without potentially achieving a “best in class” algorithm or with specific regret guarantees. This allows for the use of more “powerful” black-box machine learning algorithms to be used in these settings for which formal regret guarantees don’t exist or are hard to prove.

## Related Literature

This paper can be related to two distinct fields of study. The first is policy learning, which has gained interest from econometricians, statisticians, and computer scientists. In this line of literature, statistical methods have been developed to find the optimal policy, often from a pre-defined class of allowed policies. We can further divide these methods into model-based methods and direct-search methods.

Belonging to model-based approaches, from the computer science literature, we have Q-learning (Qian and Murphy, 2011) and A-learning (Shi et al., 2018), which both estimate a conditional expectations function (or contrast function) and then determine the optimal policy from these predictions. A short coming of this approach is that it relies heavily on correct functional form assumptions.

Among direct search methods, outcome weighting (Zhao et al., 2012) attempts to learn the optimal policy non-parametrically using an inverse probability weighting estimator (IPWE). %%What economic jargon is used here?%% However, it is well known that there is the potential for instability caused by extreme propensity scores and model specification.

To increase stability, some methods use the augmented IPWE (Zhang et al., 2012a,b; Zhao et al., 2019; Athey and Wager, 2021; Pan and Zhao, 2021), which has the double-robustness property. Athey and Wager’s (2021)’s seminal method comes with minimal optimal regret guarantees under suitable regularity conditions of the estimators. These methods come with strong theoretical guarantees but still suffer extreme weights. This problem is made more server when we only have small sample sizes.

Similar to Athey and Wager (2021), we propose using generalised double machine learning to estimate the causal effects of a decision function, using cross-fitting and classical estimators or flexible ML for the estimation of nuisance function.

Similar phenomena have been discussed in both policy learning and causal inference literature (e.g., Zhao et al., 2019; Wu et al., 2022). From (Matching-Based Policy Learning)

The evaluation of treatment assignment rules has a long history in economics. Classic examples in economics include enrolment in government welfare programs (Dehejia, 2005), job training programs (Black et al., 2003; Frölich, 2008) and judge sentencing decisions (Bushway and Smith, 2007). This has often taken the for of the minimax regret, an approach pioneered by Manski (2004).

A short coming with the literature on optimal statistical decision rules is that rules are undetermined in complex setting such as including capisity or fairness constraints.

This research is also motivated by fair machine learning / automated decision-making.

## Treatment Effect Evaluation

## Notation

Let represent a measure of welfare, signify a given outcome, denote a set of covariates and indicate a set of potential instruments, and a treatment variable . Formally, define such that .

## Average Treatment Effect Estimation

To define welfare for evaluating algorithm decisions, we introduce potential outcomes. In the binary treatment case, the realized outcome is given by

As is common analysis of statistical decision rules, we can evaluate the performance of an algorithm can be evaluated based on the distribution of outcomes induced by the algorithm.

To start, we consider estimating the treatment effect in the partially linear regression model

> I don’t think this is quite correct.

Treatment assignment problems are related to the estimation of conditonaly average treatement effects. Assuming conditonal independence of treatment, we have

Under utilitarian welfare (with normalised to utilis),, the welfare of a treatment algorithm is given by the below definition.

In this paper, we propose using (**chernozhukovLocallyRobustSemiparametric2022?**)’s method, which is a general GMM framework for estimating treatment effect-like parameters based on debiased machine learning. In the basic algorithm, the inference is based upon a method of moments estimator for some low dimensional parameter, such as the ATE, based upon the empirical analogue of the moment condition

is the score, denotes the true value of the parameter of interest and being the nuisance paramters with true values . The expectation is taken over .

The score function has the important property that

identifies when (), as well as having the Neyman orthogonality condition

where detnoes the pathwise or Gateaux derivative operator. This amounts to the scoring function being locally insensitive to perturbations around the true value of the nuisance parameters.

Validity of the approach depends on defining an appropriate score function that satisfies the above condition, with some being proposed below.

### Neyman Orthogoal scores for the linear regression model

A suitable score for the paritally linear regression model is given by

where , with true value . Here and are square-integrable functions mapping the support of to , whose true values are the following conditional expecatation functions

### Interactive Regression Model

For estimation of the average treatment effect in the interactive reegression model, we can use the following score function

where

This similar to the well-known doubly-robust estimator. It is known as doubly robust due to having the property that either the propentisy score or the outcome needs to be properly modelled.

Here is a square-integratable function mapping the support of to , and maps the support of to , where . The true values are given by

### Assumption Learners

To get consistent estimates, we need to assume sufficiently quality of the models that we use. For estimated nuciance components , sufficient condition being

Choice of the base estimators depends on the situation at hand. For example, believing the data to be sparce, we might choose to use LASSO with a dictionary of transformations on . How we choose learners is still an active area of research with not much said about choosing the theoretically best learner. That is why we advocate ensembles or what has been called “super-learning”.

The performance of the ensemble is theoretically no worse that the best performing model in used our base learners. We can therefore combine powerful machine learning estimators, which may be have good finite sample performance, with various non-parametric estimators (which are known to satisfy the conditions asymptotically).

Under the following an appropriate score function, assumptions (X) on the convergence of the learners and the use of sample-splitting, the approach works for estimating average treatment effects.

## The Double machine learning algorithm

Under the above assumpitons, assume we have a sample modelled as i.i.d coipies of random variable , hows law is determined by the probability measure . The DML algorithm dives unbiased estimates of the paramter .

**DML Algorithm** Inputs: , the Neyman orthogonal score/moment function that identifies the paramter of interest, and estimation methods for .

Take a random partition of observation indicies such that each fold is approximiatly the same size For do end do Let (The moment equation estimator) solution of

The asymptotic variance is given by

where and .

Chernozhukov et al (book) recommend using 4-5 for a medium sized data-set. For smaller data-sets, we may choose to do more splits.

Under strong identification assumption, we identify the . That is we have and that has singular values that are bounded away from zero. This method also has the beneficial quality of being Neyman orthogonal.

In most settings, our nuisance parameters will be a regression function

Chernozhukov (Book) recommend the following method for choosing the parameter. Select different poteintal ML or statistical estimators - For each , compute cross-fitted MSPE

- Select model which has the lowest MPSE - Use method as the learer of in the generic DML algorithm.

While note giving us direct individual treatment effects, we can use this on our way to building estimates of the utility of an algorithm.

## From Average to Conditional Treatment Effects

We express the CATE as the conditional expectation of an unbiased signal

where the unbasied signal takes the from

with nuisance parameters and

and

> This is from chapter 14 CausalML Book. pg 378

We can then form (I think) our individual score as

where .

Where does this leave me. I knew that it should already be unbiased. And if it wasn’t then It wouldn’t work anyway. What I think is a bit of a disapointnment is that the asympototics are already worked out. But there is lots to say about asympototics

It is unlikely that the finite sample properties are tractable. Could the bootstrap thus prove to be better in finite settings.

You are referred to **X** for a detailed treatment.

A probabilistic decision algorithm returns the probability of making a positive decision (e.g. to treat or not to treat). As part of assessing the quality of an algorithm, we want to create inference on the value of algorithm .

Given an algorithm , I define the value of the algorithm as the average value if we follow ’s decisions vs treating no one in our population:

As we have shown, the CATE is identified by (Theorem 5.2.1). This implies our value of our algorithm is given by

I will next present an algorithm for estimating this, and then show that this provides an biased estimate of the value of the algorithm.

Theorem (10.4)

We next show how our proposed bootstrap procedure can be used to efficiently compare the effect of different algorithmic policies.

## Proposed Approach

## Outline of testing procedure

We have access to an independent and identically distributed set of variables drawn from an unknown distribution . An algorithm maps covariate vectors to a decision in set . We define the value of an algorithm as

**Definition** (Welfare Improvability). Fix some . We say an algorithm consititutes a -welfare improvement if and only if

In the testing procedure, the null hypothesis is given by

with the alternative being that such an algorithm exists.

The analysis must then define a selection rule or mapping from data samples to an algorithm in :

where is the set of all finite samples of observations.

A sample-splitting procedure is then used to test for welfare improvability. First, choose the number of splits , and for each split in the data, choose the size of a hold-out set . We then perform the following three steps For 1. Split the sample into train and test sets. The training sample has observations selected uniformly at radnom. contains the remaining observations. 2. Find a candiate algorithm using . Using the selection rule, produce a candidate algorithm . 3. Test whether constitutes a -welfare improvement over .

## Test

Samples, with mean values and . Under the null hypothesis

and

We take the pooled sample . Our bootstrap sample is taken from the pooled sample by sampling from with replacement.

We generate the p-value by generating independent bootstrap samples with the p-value calculated as

### Test Asymptotic

With our ideal bootstrap, as :

Main asymptotic justification of the bootstrap, conditional on :

#### Conditional CLT for the mean

Let be iid random vectors with mean and covariance .

Delta method If is continuously differentiable, then and conditionally, then and conditionally.

Though Edgeworth expansion (a refinement of the central limit theorem) that the bootstrap has a faster convergence rate than simple normal approximations.

#### Iterated Bootstrap

If chosen correctly, the iterated bootstrap can have higher rate of convergence than the non-iterated bootstrap.

Double Bootstrap Idea: Use an iterated version of the bootstrap to correct the p-value (the bootstrap does not guarantee that the p-value will be distributed uniformly under the null, although it should). Let be the p-value based on . Let be the random variable obtained by resampling from . We get the adjusted p-value from

**Set-up** Generate from the fitted null distribution (I think this is the pooled sample), calculating the test statistic . We then fit the null distribution to obtaining for : - Generate from - Calculate the test statistic from them Let . Let

## Testing Groups

We assume we have an additional binary variable indicating group member ship with .

**Definition** (Group Welfare Improvability) Fix a class of algorithms and a tuple . We say the algorithm constitutes a -improvable within class if there exists an algorithm such that

We make adjustments for multiple testing.

## Main Results

#### Assumption

A1: Independently and Identically Distributed Data The observations are independent and identically distributed, random vectors or scalars

A2: Bounded and Measurable Treatment Policies - The treatment policies and and reasurable functions satisfying for all and . - The difference is uniformly bounded implying there exists an such that for all .

A3: Finite Second Moments - The conditional treatment effect satisfies - The product has finite variance:

A4: Consistent Estimation of Treatment Effects The estimated treatment effects

where the estimation errors satisfy - - is independent of due to cross-fitting - for some and constant

A5: Cross-Fitting and Independence The estimation of is performed using cross-fitting, ensuring that is approximately independent of .

A6: Regularity Conditions for the Bootstrap - The samples are generated by resampling the pairs with replacement - The bootstrap replicates mimic the sampling variability of under the null hypothesis

A7: Lindeberg-Feller Conditions For the sequence defined by , the Lindeberg condition holds:

A8: Non-degeneracy The variance .

As above, we define the test statistic

**Theorem** (Asymptotic Distribution of the Test Statistic). Under assumptions A1-A4, A7, and A8, we have

where .

**Proof** By A4 and the law of iterated expectations:

Since and are independent (due to cross-fitting), and , we have

By A4, is .

Finally, under assumptions A1, A2, A3, A7, and A8, the Lindeberg-Feller CLT applies to , yielding

Furthermore, for the term involving :

END PROOF

**Theorem** (Consistency of Bootstrap) Under Assumptions A1-A8, the bootstrap p-value for testing the null hypothesis

using test statistic converges in probability to the true p-value as . That is

where with .

Proof Define the bootstrap test statistic as

Our first want to show that, conditional on the observed data:

where and .

Given the data, the bootstrap sample consists of i.i.d. draws from the empirical distribution of . Conditional on the data, the bootstrap sample satisfies the Lindeberg condition (since the original sample does). Therefore,

Furthermore, under A2 and A3, .

Combining the above, the bootstrap distribution of converges in distribution (conditional on the data) to the same normal distribution as .

We now turn to showing convergence of the bootstrap p-value. We defined the p-value as

where denotes the probability under the bootstrap distribution.

## Technical Considerations

Super Learning / Ensemble learning - convex approach and best learner. - Use a mix of traditional methods with known convergence rates, as well as LM methods

## Simulations

We consider the following DGPs ### Linear DGP with Simple Threshold Decision Rules

We first consider a single variable distributed uniformly on 0 and 1, where our two algorithms are simple threshold decision rules

The true utility utility is given by . We assume our DDML estimates are unbiased and given by

The decision rules are given by

We control the true difference in algorithm welfare () by adjusting and . We also control estimation variance by changing .

The expected utility is given by

### Nonlinear DGP with Quadratic Decision Rules

We next use the following DGP

The decision algorithms are again threshold rules

The expected difference in utility between the two algorithms is

### Binary Outcome DGP with Logistic Decision Rules

• **Independent Variable:**

• **True Utility:**

• **Estimated Utility:**

The expected difference in utility between the two algorithms is

## Empirical Application

In our first empirical example, we utilise data from the randomised controlled trial (RCT) conducted by Finkelstein et al. (2020) on the Camden Coalition of Healthcare Providers’ “hotspotting” program. The program targeted “superutilizers”—patients with exceptionally high healthcare usage—and aimed to reduce hospital readmission rates by employing a care-transition model involving multidisciplinary teams of nurses, social workers, and community health workers. The intervention group consisted of 800 hospitalised patients with medically and socially complex conditions, all of whom had experienced at least one additional hospitalisation in the preceding six months.

The critical variables in the study include hospital readmission within 180 days and the costs associated with readmission and other health-related expenditures. The data collected included patient demographics, hospitalisation history, and post-discharge interactions with care teams. You can see a complete list of variables used in the appendix below. Despite the widespread optimism surrounding hotspotting, the RCT results indicated no significant difference in average treatment effect of readmission rates between the intervention and control groups, challenging prior observational claims of the program’s efficacy.

We apply this data to evaluate the impact of algorithmic decision-making in healthcare settings. Specifically, we investigate heterogeneity in treatment effects and compare two automated algorithms in their ability to identify high-risk patients who could benefit most from targeted interventions.

We find substantial variation in the treatment effects estimated through double machine learning. Further, sensitivity tests are performed in the appendix.

## Conclusion

## Alternative - Split Sample

We consider the training set and a test point sampled i.i.d. from some unknown distribution .