# PARETO-SMOOTHED INVERSE PROPENSITY WEIGHING FOR CAUSAL INFERENCE

FUJIN ZHU[1, 2], ADI LIN[2], GUANGQUAN ZHANG[2], JIE LU[2], DONGHUA ZHU[1]

*[1]School of Management and Economics, Beijing Institute of Technology*
*Beijing 100081, China*

*[2]Center for Artificial Intelligence (CAI), School of Software, FEIT, University of Technology Sydney, NSW 2007, Australia*

Causal inference has received great attention across different fields ranging from economics, statistics, biology, medicine, to machine learning. Observational causal inference is challenging because confounding variables may influence both the treatment and outcome. Propensity score based methods are theoretically able to handle this confounding bias problem. However, in practice, propensity score estimation is subject to extreme values, leading to small effective sample size and making the estimators unstable or even misleading. Two strategies– truncation and normalization – are usually adopted to address this problem. In this paper, we propose a new Pareto-smoothing strategy to tackle this problem. Simulations and a real-world example validate the effectiveness.

## 1. Introduction

To minimize the confounding bias in observational causal inference, statistical "case-mix adjustment" techniques are frequently adopted. Among them, Rosenbaum and Rubin [1] introduced the propensity score to summarize the information required to control the confounders. The propensity score is the conditional probability of an individual to be assigned to the treatment group. Theoretically, one can account the difference between the treatment and control groups by directly modelling the assignment mechanism with propensity scores, and thus making the treated and control populations more comparable.

Though propensity score provides us a convenient solution to ease the issue of confounding, the true propensity scores are intrinsically unknown in pure observational studies. A practical concern is that the causal effect may be difficult to estimate precisely if the estimated propensity score is close to zero for a substantial fraction of the population [2]. This is a particular concern in setting with many covariates or the assignment mechanism is highly skewed.

When many of the estimated propensity scores are close the zero, the distribution of their reciprocals – the inverse propensity (IP) weights – can have a heavy right tail, which will lead to unstable inverse propensity weighting estimates, sometimes with infinite variance. To cope with this problem, methods including truncation and self-normalization have been proposed [3-5]. In this paper, we propose a new Pareto-smoothing strategy. Compared with truncation, our method is less biased. Compared with the normalization strategy, our experiment result shows that they both converge to the true value if we have enough data. One special merit of our method is that it is more stable in the small sample size cases, which are common in many real problems.

The reminder of the paper is organized as follows. In Section 2, we formalize the causal inference problem, introduce the concept of propensity score and two stabilization strategies for propensity score based estimators. Section 3 illustrates the proposed strategy and methods for parameter estimation. Experiments on simulated and real data are conducted in Section 4. Section 5 concludes the paper.

## 2. Causal Inference and Inverse Propensity Weighting

### 2.1. *Notation and Problem Formalization*

Suppose there are $N$ units $X_i$ $(i = 1, \dots, N)$, denote the treatment condition for unit $i$ with $A_i$, where $A_i = 0$ indicating that unit $i$ received the control treatment and $A_i = 1$ the active treatment. Let $Y$ be the outcome variable of interest. $Y_i(A)$ is defined as the potential outcome of unit $i$ had she received treatment $A$. We postulate the existence of a pair of *potential outcomes* for each unit, $\big(Y_i(0), Y_i(1)\big)$, and the observed outcome $Y_i = Y_i(A_i) = A_i Y_i(1) + (1 - A_i)Y_i(0)$. With this notation, the individual treatment effect for unit $i$ is $\tau_i = Y_i(1) - Y_i(0)$ and the average causal effect (aka, average treatment effect, ATE) is its expectation, i.e., $\tau = \mathbb{E}[\tau_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$.

ATE measures the expected causal difference of a population if all of them were treated versus all were untreated, which is generally different from the conditional difference $\mathbb{E}[Y_i|A_i = 1] - \mathbb{E}[Y_i|A_i = 0]$. As a baseline, we also denote the empirical conditional difference as the naïve ATE estimator in Eq. (1)

$$\hat{\tau}_{naive} = \frac{1}{N_1}\sum_{i=1}^{N} A_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - A_i)Y_i \qquad (1)$$

where $N_1 = \sum_{i=1}^{N} A_i$ is the number of treated and $N_0 = N - N_1$ the number of control.

Estimating ATE from observational data is generally impossible because of the fundamental problem of causal inference [4]. Under the conditional exchangeability (or unconfoundedness) condition, $Y_i(0), Y_i(1) \perp\!\!\!\perp A_i|X_i$, Pearl [6] proves that the ATE can consistently estimated by Eq. (2) as:

$$\tau = \int (\mathbb{E}[Y_i|A_i = 1, X_i = x] - \mathbb{E}[Y_i|A_i = 1, X_i = x])dP(x) \tag{2}$$

This formula is also called the G-computation formula [7] and the back-door adjustment formula [6]. Although feasible for estimating ATE in principle, it is in practice infeasible to implement with many covariates. In the following section, we introduce the propensity score and its importance for solving this challenge.

### 2.2. *Propensity Score and Inverse Propensity Weighting (IPW)*

As discussed earlier, adjusting for all observed covariates to eliminate confounding bias may go out of the question. As the coarsest balancing score [4], the propensity score is a scalar proxy of them that suffices for removing the bias associated with imbalance in the pre-treatment covariates and is defined as:

**Definition 1 (Propensity Score, PS)** *The propensity score* $e(X_i)$ *is the conditional probability of an individual* $X_i$ *to be assigned to the treatment group.*

Defining the inverse propensity weight (IP weight) for unit $i$ as

$$w_i = \frac{1}{p(A_i|X_i)} = \frac{\mathbb{1}(A_i = 1)}{e(X_i)} + \frac{\mathbb{1}(A_i = 0)}{1 - e(X_i)} \tag{3}$$

where $\mathbb{1}(a_i = a)$ is the indicator function, we can build a balanced pseudo-population where the treatment assignments are randomized and all confounding is removed. The conditional difference in this super population consistently estimates $\tau$ by the inverse propensity weighted (IPW) estimator [8] [1]

$$\hat{\tau}_{IPW} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}(A_i = 1)}{p(A_i = 1|X_i)} Y_i - \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}(A_i = 0)}{p(A_i = 0|X_i)} Y_i \tag{4}$$

Note that the propensity scores $e_i = e(X_i)$ occur in the denominator of Eq. (3), we thus need to make the "positivity" or "overlapping" assumption, for all $i, 0 < e(X_i) < 1$, so that the IP weights are bounded, $w_i < \infty$. Theoretically, $\hat{\tau}_{IPW}$ is unbiased and consistent under this positivity assumption if we have infinite many observations. However, for finite data, the estimated propensities $\hat{e}(X_i)$ can be very close to zero for some $X_i =$ x. An extreme case may occur that there are regions of covariate values observed in only one of the two treatment conditions. In this case, the IP weights $w_i$ will be highly variable and even unbounded, thus estimation based on then will be unstable and misleading.

### 2.3. *Stabilization by Truncation and Normalization*

To remedy the issue of high variability, there are mainly two strategies for stabilization [5]: truncation (aka clipping) and normalization of the propensity score. The truncated IPW estimator for causal inference is given by

$$\hat{\tau}_{T-IPW} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(A_i = 1)}{g_i(A_i|X_i)}Y_i - \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(A_i = 0)}{g_i(A_i|X_i)}Y_i \qquad (5)$$

with the estimated treatment probabilities truncated by a constant $C$:

$$g_i(A_i|X_i) = \begin{cases} C, \quad constant \quad if\ p(A_i|X_i) < \delta \\ p(A_i|X_i), \qquad\qquad else \end{cases} \qquad (6)$$

A consequence of PS truncation is the introduction of bias in the estimated PS, which in turn causes bias in PS-based causal estimators. Moreover, the cut-point $\delta$ is usually unknown and choosing it relies on experience or intuition. Recently, [9] propose a data-adaptive PS truncation algorithm which can select the optimal truncation threshold adaptively, but it is specially designed for target maximum likelihood estimators [10].

Alternatively, the normalized IPW estimator [5,11] divides the IP weights by the empirical mean of each treatment group and is given by

$$\hat{\tau}_{N-IPW} = \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(A_i = 1)}{p(A_i = 1|X_i)}Y_i}{\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(A_i = 1)}{p(A_i = 1|X_i)}} - \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(A_i = 0)}{p(A_i = 0|X_i)}Y_i}{\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(A_i = 0)}{p(A_i = 0|X_i)}} \qquad (7)$$

For a set of IP weights $W = \{w_1, w_2, \dots, w_N\}$, denote $\overline{W} = \frac{1}{N}\sum_{n=1}^{N} w_n$ and $\overline{W^2} = \frac{1}{N}\sum_{n=1}^{N} w_n^2$, we also define the effective sample size as $N_{eff} = \frac{N\overline{W}^2}{\overline{W^2}}$, which will be used as a measure of stability in the experiment sections. If the weights are highly imbalanced, they will have a high sampling variance, and the resulting estimate will be unreliable with a very small $N_{\text{eff}}$.

## 3. Pareto Smoothing for Causal Inference

Our method builds upon results in the extreme value theory [12]. The idea is simple, given the estimated IP weights $\{w_1, w_2, \dots, w_N\}$, we fit a generalized Pareto distribution (GPD) on these extreme values, and replace them with order statistics of the fitted GPD. By this smoothing strategy, we try to stabilize the IP weights while keep the information of their relative order.

### 3.1. *The Generalized Pareto Distribution*

Among the series of extreme value distributions in the extreme value theory [12], the generalized Pareto distribution, named by Pickands [13], is a family of extreme value distributions that is often used to model the tails of another distribution. A GPD is specified by the location $\mu$, scale $\sigma > 0$, and shape $\kappa$:

$$F(x) = \left[1 - \left(1 + \frac{\kappa(x - \mu)}{\sigma}\right)^{-\frac{1}{\kappa}}\right]\mathbb{1}(\kappa \neq 0) + \left(1 - e^{-\frac{x-\mu}{\sigma}}\right)\mathbb{1}(\kappa = 0) \tag{8}$$

where the $\mu$ is a lower bound, i.e., $x \in (\mu, \infty)$. Pickands [13] proves that if an unknown distribution function $F(x)$ lies in the “domain of attraction” of some extremal distribution function, then $F(x)$ has a generalized Pareto upper tail.

### 3.2. *Parameters Estimation*

To fit the parameters $\boldsymbol{\theta} = (\mu, \sigma, \kappa)$, we follow [14] and choose the location parameter $\mu$ so that the size of the *upper-tail* is

$$M = min\left(\lfloor 0.2S \rfloor, \lfloor 3\sqrt{S} \rfloor\right) \tag{9}$$

Having decided the location $\mu$, the other two parameters $\sigma$ and $\kappa$ can be estimated by maximum likelihood [12]. Given a random sample $X = \{x_1, x_2, \dots, x_M\}$, [15] reparametrize Eq. (8) by two parameters $(\alpha, \kappa)$, where $\alpha = \kappa/\sigma$, and the estimate $\hat{\alpha}$ is obtained by maximizing a *profile likelihood function* with a weakly informative prior, $\kappa$ and $\sigma$ are estimated by

$$\hat{\kappa} = \frac{1}{M}\sum_{i=1}^{M} log(1 - \hat{\alpha}x_i), \qquad \hat{\sigma} = \frac{\hat{\kappa}}{\hat{\alpha}} \tag{10}$$

### 3.3. *Summary of the Pareto-smoothed IPW Estimator*

Given a set of $N$ observations $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^{N}$, our proposed Pareto-smoothed IPW method can be easily implemented and proceeds as follows:

1. Estimate the propensity scores and get $\{e_i, i = 1,2, \dots, N\}$;
2. Sort $e_i$ descending, calculate $M$ by Eq. (9) and choose the corresponding $\mu$;
3. Let $\mu = 1/\mu$, and calculate the IP weights $\{w_i, i = 1,2, \dots, N\}$;
4. Estimate $\sigma$ and $\kappa$ using the largest $M$ IP weights by Eq. (10);
5. Replace the largest $M$ weights with ordered statistics of the fitted GPD, and obtain the “Pareto-smoothed” weights $\{w_i^{PS}, i = 1,2, \dots, N\}$;
6. Estimate the ATE using $\{w_i^{PS}, i = 1,2, \dots, N\}$ by

$$\hat{\tau}_{PS-IPW} = \frac{1}{N}\left(\sum_{i:A_i=1} \frac{w_i^{PS}}{\overline{w^{PS}}_t} Y_i - \sum_{i:A_i=0} \frac{1-w_i^{PS}}{\overline{w^{PS}}_c} Y_i\right) \tag{11}$$

where $\overline{w^{PS}}_t = \frac{1}{N}\sum_{i:A_i=1} w_i^{PS}$ and $\overline{w^{PS}}_c = \frac{1}{N}\sum_{i:A_i=0}(1 - w_i^{PS})$.

## 4. Experimental Study

In this section, we validate our proposed method using simulated and semi-simulated data. In all the experiments, we use logistic regression to fit the propensity score model. The mean absolute error (MAE) $\epsilon_{ATE} = \frac{1}{n}|\sum_{i=1}^{n}(Y_i(1) - Y_i(0) - \hat{\tau}_i)| = \frac{1}{n}|\sum_{i=1}^{n}(\tau_i - \hat{\tau}_i)|$ will be reported. An application on a real world job training study is also conducted.

### 4.1. *Simulated and Semi-simulated Data*

The specific data-generating process of our simulation is: $X_{i,1} \sim Bernoulli(0.5)$, $X_{i,2} \sim Binomial(3, 0.5)$, $(A_i|X_i) \sim Bernoulli(\text{Sigmoid}(-1.3 - 3X_{i,1} + 3X_{i,2}))$, $(Y_i|X_i, A_i) \sim Bernoulli(\text{Sigmoid}(-2 - 2X_{i,1} + 3X_{i,2} + 3A_i + 2AX_i))$. We simulate data with sample size $N$ ranging from $100$ to $10^5$, and run each simulation 10 times. Comparisons of the MAE and effective sample size are in Fig. 1. We known that on one hand, $\hat{\tau}_{PS-IPW}$ is less biased than $\hat{\tau}_{T-IPW}$. On the other hand, the estimate of $\hat{\tau}_{PS-IPW}$ converges together with $\hat{\tau}_{IPW}$ and $\hat{\tau}_{N-IPW}$ to the true estimate as the sample size gets large, say$10^4$. Actually, both $\hat{\tau}_{IPW}$ and $\hat{\tau}_{N-IPW}$ are theoretically unbiased, but when the sample size is relatively small, their estimates are unstable compared with our Pareto-smoothed estimator. This indicates the advantage of our method in the small data cases. As to the effective sample size, since many of the IP weights are truncated to the same value, the effective sample size of $\hat{\tau}_{T-IPW}$ is supposed to be high. However, $\hat{\tau}_{PS-IPW}$ has higher effective sample size than $\hat{\tau}_{IPW}$ and $\hat{\tau}_{N-IPW}$ in general.

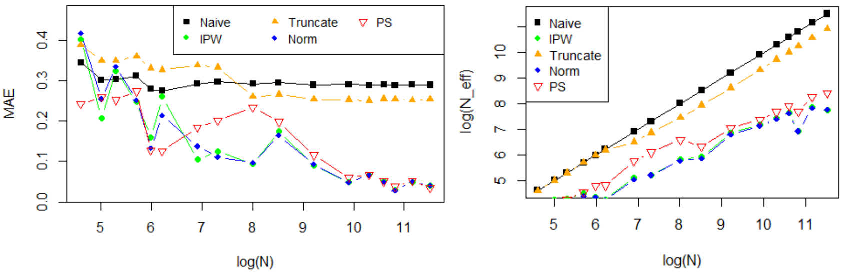

Figure 1. Comparison of the MAE (left) and log effective sample sizes (right) of different estimators.

Table 1. ATE estimates and effective sample size for the IHDP data.

| | Naïve | IPW | T-IPW | N-IPW | PS-IPW |
|---|---|---|---|---|---|
| MAE | 4.782 | 0.32 | 2.894 | 0.008 | **0.0008** |
| $N_{\mathrm{eff}}$ | 747 | 304.247 | 608.234 | 292.390 | 273.241 |

We also evaluated the performance of our algorithm through the semi-simulated IHDP dataset introduced in [16]. It is based on covariates from a real randomized experiment that evaluated the impact of the IHDP on the subjects' IQ test scores at the age of three while all outcomes are simulated. In total, the dataset consists of 747 subjects (139 treated, 608 control), and 25 covariates measuring properties of children and their mothers. The MAE and effective sample size results are listed in Table 1. Our proposed method outperforms other estimators regarding MAE. Actually, while the truncation strategy suffers a relatively high bias, the performances of $\hat{\tau}_{N-IPW}$ and $\hat{\tau}_{PS-IPW}$ are very close.

### 4.2. *Real Data: NSW Job Training Study*

As an application of the methods introduced in this paper, we use the randomized experiment data of [17], which is part of the "National support work" (NSW) demonstration programme implemented in the mid-1970s to study whether a systematic job-training programme would increase post-intervention income levels among workers [18]. In this paper, we simply use the `nsw` dataset in the R package ATE[a], which provides LaLonde's original 722 observations (297 treated and 425 control). The kernel density fits of the estimated IP weights in Fig. 2 indicates the imbalance between the treatment and control group. The resulting estimates are $\hat{\tau}_{naive} = -537.803, \hat{\tau}_{IPW} = 3.696, \hat{\tau}_{T-IPW} = 736.033, \hat{\tau}_{N-IPW} = 798.488,$ and $\hat{\tau}_{PS-IPW} = 805.881$. The result again validate the performance similarity between our Pareto-smoothing strategy and the normalization strategy.

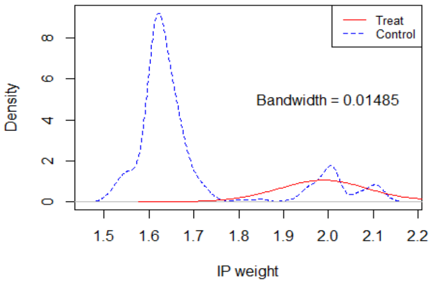

Figure 2. Comparison of density distribution of the estimated IP weights for the NSW dataset.

[a] https://cran.r-project.org/package=ATE

## 5. Conclusion

In this paper, we concluded two stabilization strategies for handling the problem of IP weights variability in PS-based causal inference, and proposed a new Pareto-smoothing strategy. Empirical results indicate that the proposed method has appealing advantages, i.e., it is less biased than brute-force truncation and more stable than the normalization strategy in the small sample size setting. Though empirically appealing, our future work will be in its theoretical analysis as well as its applications in other causal effect estimators, for example, propensity score matching and balancing estimators.

## References

1. P. R. Rosenbaum and D. B. Rubin, Biometrika **70** (1), 41 (1983).
2. S. Athey, G. Imbens, T. Pham, and S. Wager, arXiv preprint arXiv:1702.01250 (2017).
3. S. L. Morgan and C. Winship, *Counterfactuals and causal inference*. (Cambridge University Press, 2014).
4. G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. (Cambridge University Press, 2015).
5. M. A. Hernán and J. M. Robins, *Causal Inference*. (Boca Raton: Chapman & Hall/CRC, forthcoming, 2018).
6. J. Pearl, *Causality: Models, Reasoning and Inference*. (Cambridge University Press, 2000).
7. J. Robins, Mathematical modelling **7** (9-12), 1393 (1986); J. M. Robins, in *Latent variable modeling and applications to causality* (Springer, 1997), pp. 69.
8. D. G. Horvitz and D. J. Thompson, Journal of the American statistical Association **47** (260), 663 (1952).
9. C. Ju, J. Schwab, and M. J. van der Laan, arXiv preprint arXiv:1707.05861 (2017).
10. M. S. Schuler and S. Rose, American journal of epidemiology **185** (1), 65 (2017).
11. A. Swaminathan and T. Joachims, in *NIPS* (2015), pp. 3231.
12. S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*. (Springer, 2001).
13. J. Pickands Iii, the Annals of Statistics, 119 (1975).
14. A. Vehtari, A. Gelman, and J. Gabry, arXiv preprint arXiv:1507.02646(2015).
15. J. Zhang and M. A. Stephens, Technometrics **51** (3), 316 (2009).
16. J. L. Hill, Journal of Computational and Graphical Statistics **20** (1), 217 (2011).
17. R. J. LaLonde, The American economic review, 604 (1986).
18. K. C. G. Chan, S. C. P. Yam, and Z. Zhang, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **78** (3), 673 (2016).