
Estimation of Treatment Effects in Extreme and Unobserved Data

Jiyuan Tan*

Department of Management Science and Engineering
Stanford University
jiyuantan@stanford.edu

Jose Blanchet[†]

Department of Management Science and Engineering
Stanford University
jose.blanchet@stanford.edu

Vasilis Syrgkanis[‡]

Department of Management Science and Engineering
Stanford University
vsyrgk@stanford.edu

Abstract

Causal effect estimation seeks to determine the impact of an intervention from observational data. However, the existing causal inference literature primarily addresses treatment effects on frequently occurring events. But what if we are interested in estimating the effects of a policy intervention whose benefits, while potentially important, can only be observed and measured in rare yet impactful events, such as extreme climate events? The standard causal inference methodology is not designed for this type of inference since the events of interest may be scarce in the observed data and some degree of extrapolation is necessary. Extreme Value Theory (EVT) provides methodologies for analyzing statistical phenomena in such extreme regimes. We introduce a novel framework for assessing treatment effects in extreme data to capture the causal effect at the occurrence of rare events of interest. In particular, we employ the theory of multivariate regular variation to model extremities. We develop a consistent estimator for extreme treatment effects and present a rigorous non-asymptotic analysis of its performance. We illustrate the performance of our estimator using both synthetic and semi-synthetic data.

1 Introduction

We are interested in studying the effect of treatment e.g., different policies and drugs, on rare yet impactful events such as large wildfires, hurricanes, tsunamis and climate change. These kinds of events happen at an extremely low frequency, but they can cause considerable damage to properties and pose serious threats to people's lives. For instance, we may want to know the effect of more

*Jiyuan Tan was partially supported by NSF Award IIS-2337916.

[†]Jose Blanchet gratefully acknowledges support from DoD through the grants Air Force Office of Scientific Research under award number FA9550-20-1-0397 and ONR 1398311, also support from NSF via grants 2229012, 2312204, 2403007 is gratefully acknowledged.

[‡]Vasilis Syrgkanis was supported by NSF Award IIS-2337916.

infrastructure investment or other kinds of precautionary policies on earthquakes. In many applications – from financial risk to environmental policy – it isn't enough to know how a treatment changes the average outcome; decision-makers care about whether it alters the extreme tail. More formally, we may want to estimate the effect of treatment D on outcome Y , conditioning on some extreme events. Estimating this kind of effect can help policymakers evaluate the impact of a policy and choose the best policy to reduce economic loss and save more lives when disasters happen.

Despite its clear importance, existing methods fall into two largely disconnected strands, each of which cannot fully address this question. One approach comes from the causal inference literature. Causal inference provides a comprehensive framework for counterfactual reasoning. Causal effect estimation is an important problem in this area, which finds wide applications in healthcare, education, business decision-making, and policy evaluation. Classic causal inference literature mainly focuses on estimating the average effects among certain groups. Little attention is paid to the causal effect on rare events. The scarcity of extreme data makes inference more challenging than in classic settings. As a result, naively applying classic causal inference estimation methods will produce poor results with large statistical error. For example, when making policies about earthquakes, we are usually unable to see a strong signal from historical data, as large earthquakes rarely occur and there are fewer samples in the dataset.

On the other hand, the Extreme Value Theory (EVT) studies the tail behaviors for statistical distributions, which provides the ideal tools for analyzing rare events. However, this approach does not take the data structure into consideration. In particular, it does not accommodate counterfactual treatments or adjust for covariates, so it cannot tell us what would happen under an intervention.

To bridge these gaps, we combine causal inference theory with EVT to provide a novel framework for extreme effect measurement. Following researches in EVT Coles et al. [2001], we use a multivariate regularly varying variable U to model extreme events. The rare event can be modeled by the event $\{\|U\| > t\}$ for large t . Our proposed estimand can be viewed as the Average Treatment Effect (ATE) conditioning on $\{\|U\| > t\}$ with rescaling as t increases to infinity. Detailed definition and explanation can be found in Section 3. Estimation is challenging because the limiting tail distribution is unknown and must be inferred from finite samples. To improve data efficiency and inference accuracy, we combine tail observations with moderate-frequency data in an extrapolation scheme, leveraging EVT insights alongside causal-inference techniques to achieve efficient estimation.

To the best of our knowledge, we are not aware of any work in the literature that considers this problem. In this paper, we take the first step to measure the treatment effect on extreme events. To be more specific, our contributions can be summarized as follows.

1. We propose a measure for the treatment effect on rare events named Normalized Extreme Treatment Effect (NETE), which essentially measures the magnitude of treatment on tailed events.
2. We develop two consistent estimators for NETE—a doubly robust (DR) estimator and an inverse propensity weighting (IPW) estimator—by combining recent advances in multivariate tail-dependence estimation Zhang et al. [2023] with double machine learning methodology Chernozhukov et al. [2018], and derive finite-sample, non-asymptotic error bounds.
3. Synthetic and semi-synthetic experiments demonstrate a good practical performance of our proposed estimator as compared to baseline estimators adapted from standard causal inference literature.

Related Work We briefly review some relevant literature in EVT and causal inference. Coles et al. [2001] provides a comprehensive introduction to EVT. A large amount of work focuses on the univariate setting Davison and Smith [1990], Leadbetter [1991], Pickands III [1975], Smith [1989]. Recently, there have been many recent works on the multivariate generalization of these results Avella-Medina et al. [2022], Zhang et al. [2023]. Causal effect estimation is a classical problem in causal inference [Rubin, 1974]. Common estimators include IPW [Rosenbaum and Rubin, 1983], DR methods [Bang and Robins, 2005, Kang and Schafer, 2007, Chernozhukov et al., 2016, 2017, 2018], Targeted Maximum Likelihood Estimation (TMLE) [van der Laan and Rubin, 2006]. There have been some efforts in the literature trying to combine the two research areas. Gissibl and Klüppelberg [2018] considers a special kind of Structural Causal Model (SCM) and shows that the proposed SCM is a kind of max-linear model. They also analyze the asymptotic distribution of their model. Chernozhukov and Du [2006], Zhang [2018], Deuber et al. [2024] consider the task of estimating the

extreme Quantile Treatment Effect (QTE). Aloui et al. [2023], Bodik [2024] assume the outcome falls into the domain of attraction and define the extreme treatment effect as the difference between EVI. Huang et al. [2024] study the estimation of extreme quantile effect and extreme average treatment effect in the continuous treatment setting. The other line of work Gnecco et al. [2021], Mhalla et al. [2020], Bodik et al. [2023] uses EVT to help causal discovery. However, we want to point out that the problems these works consider are quite different from our setting. The most similar setting would be extreme QTE estimation Chernozhukov and Du [2006], Zhang [2018], Deuber et al. [2024], but the QTE still cannot capture on how the expectation of the outcome changes under intervention. While all the previous work models extreme events using univariate variables, we model it using multivariate regularly varying variables. Moreover, our work can also be placed in the broader context in the literature that studies causal effects of treatment beyond the mean, for example, distributional effect Abadie [2002], Gautier and Hoderlein [2011], Hohberg et al. [2020], Ratio- and Log-Ratio-Based Effects Cole and Hernán [2002], VanderWeele [2013], CATE Shalit et al. [2017], Wager and Athey [2018], Abrevaya et al. [2015].

2 Preliminary

Causal Inference. We use the potential outcome framework Rubin [1974] in this paper. Let W, D, Y be the covariate, binary treatment and outcome, respectively. We denote $Y(d)$ to be the potential outcome when the treatment is set to be d and assume consistency i.e., $Y(D) = Y$ throughout the paper. The Average Treatment Effect (ATE) is defined as

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)].$$

The ATE measures the effect of a treatment on the outcome Y . In the policy-making example, D is an indicator of whether to use the policy or not. W is a covariate that may influence D , like the geographic features of a place, which will influence the local government's decision on policies, and Y can be the economic loss. The ATE in this case provides information about how much loss can be saved if a policy is enforced. Under the following exogeneity and overlap assumptions, the ATE can be identified using the g-formula $\mathbb{E}[\mathbb{E}[Y|W, D = 1] - \mathbb{E}[Y|W, D = 0]]$.

Assumption 2.1 (Exogeneity). The data generation process satisfies $(Y(1), Y(0)) \perp D \mid W$.

Besides, the following overlap assumption is also often needed for non-asymptotic analysis.

Assumption 2.2 (Overlap). There exists constant $c_p \in (0, 1/2)$ such that the propensity score $p(w) = P(D = 1 \mid W = w) \in [c_p, 1 - c_p]$, $\forall w \in \mathcal{W}$.

This assumption ensures that there is no extremely high or low propensity, which can make estimators unstable.

Extreme Value Theory. The study of extremity is mainly concerned with the tail behaviors of heavy-tailed distributions, which are often modeled by the regularly varying distributions. In this paper, we modeled extremity by multivariate regularly varying distributions.

Definition 2.3. A random variable $U \in \mathbb{R}_+^d$ is called regularly varying with index $\beta \in (0, \infty)$ if for any norm $\|\cdot\|$ in $\mathbb{R}^d$ and positive unit sphere $\mathbb{S}^+ = \{x \in \mathbb{R}_+^d : \|x\| = 1\}$, there exists a probability measure $S(\cdot)$ on $\mathbb{S}^+$ and a sequence $b_n \rightarrow \infty$ such that $n P((\|U\|/b_n, U/\|U\|) \in \cdot) \xrightarrow{w} c_U \cdot \nu_\beta \times S$ for some constant $c_U > 0$, where $\cdot \times \cdot$ is the product measure and $\nu_\beta([r, \infty)) = r^{-\beta}$ for all $r > 0$.

The parameter $\gamma = 1/\beta$ is called the Extreme Value Index (EVI), which characterizes the decay rate of the tail. Notice that this definition implies that as $b_n \rightarrow \infty$, the norm of and $\|U\|$ and its angle $U/\|U\|$ become asymptotically independent. We will leverage this fact for estimation in later sections. A typical example of regularly varying distributions is the Pareto distribution.

Definition 2.4. The density of a Pareto (type II) distribution with index $\beta \in (0, \infty)$ is $f(x) = \beta(1+x)^{-(\beta+1)}$, $\forall x > 0$.

Definition 2.3 implies that the rescaled norm of a regularly varying variable is asymptotically a Pareto distribution.

Notations. In the rest of the paper, we use $\|\cdot\|$ and $\|\cdot\|_1$ as a shorthand for ℓ_1 -norm. We use the asymptotic order notation $o(\cdot)$, $O(\cdot)$ and $\Theta(\cdot)$. We use $\mathbb{E}[\cdot]$ to represent expectation. For a matrix

A , we denote $A_{\cdot,i}$ to be its i -th column. $\text{Unif}([a, b])$ is the uniform distribution on interval $[a, b]$ and $\text{Ber}(p)$ is the Bernoulli distribution with expectation p . $\perp$ represents the independence relationship between two random variables.

3 Treatment Effect on Extreme Events

3.1 Extreme Semi-parametric Inference

While standard causal estimands capture average effects of D on Y , they obscure what happens in the tails—i.e., when rare, high-impact events occur. To address this, we model rare events with an explicit extreme factor U . The data we consider is of the form $\{(X_i, D_i, Y_i, U_i)\}_{i=1}^N$, where D , and Y are as defined in Section 2, $W = (X, U)$ is the covariate, U is the extreme part of the covariate. We use $\|U\|$ to model the severity of rare events—large norms indicate more extreme realizations. For example, in a hurricane-loss application, U might be the vector of maximum wind speed, rainfall, and storm surge; X the region's location; D the level of infrastructure investment; and Y the resulting economic loss.

In what follows, we introduce a novel estimand that quantifies the causal effect of D on Y specifically in the tail region defined by large $\|U\|$. We then establish conditions for its identification under multivariate regular variation and propose two consistent estimators. We will make the following i.i.d. assumption.

Assumption 3.1. The random variables $\{(X_i, D_i, Y_i, U_i)\}_{i=1}^N$ are i.i.d.. Furthermore, U is regularly varying.

We are interested in the effect of treatment on the tail events of U . Similar to ATE, a naive definition of the extreme treatment effect would be

$$\theta^{\text{ETE}} = \lim_{t \rightarrow \infty} \mathbb{E}[Y(1) - Y(0) \mid \|U\| > t], \quad (3.1)$$

which is simply ATE conditioning on large $\|U\|$. However, in the case of extreme effects, the outcome may be unbounded due to the presence of extreme noise. As t increases to infinity, this effect may increase to infinity, making this quantity meaningless. Considering the climate change example, it is possible that dramatic climate change will damage or even destroy human societies, causing the effects of some policies to explode even though the policy can effectively reduce losses and slow down the process. Fortunately, regularly varying distributions have the nice property that as t increases to infinity, $\|U\|/t \mid \|U\| > t$ converges weakly to the Pareto distribution (See Definition 2.3). Inspired by this property, we can normalize the quantity $Y(1) - Y(0) \mid \|U\| > t$ by its growth rate. To characterize the growth of this quantity, we introduce the following polynomial growth assumption.

Assumption 3.2 (Asymptotic Homogeneous Property). We assume that the covariate X is bounded, i.e. $\|X\| \leq R$. Let $f(X, D, U) = \mathbb{E}[Y \mid X, D, U]$. There exists a L -Lipschitz continuous function $g(x, d, u)$ and a function $e(t) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ that satisfies $\lim_{t \rightarrow \infty} e(t) = 0$ and

$$\left| \frac{f(x, d, tu)}{t^\alpha} - g(x, d, u) \right| \leq e(t), \quad \forall x \in B_R, u \in S^{d-1}.$$

This assumption characterizes the growth of the outcome with respect to the extreme noise. In many real-world examples, this assumption is satisfied. For instance, research show that landslide volume often follows a power-law relationship with rainfall intensity Tuganishuri et al. [2024]; the economic loss caused by hurricanes scales polynomially with the maximum wind speed Zhai and Jiang [2014]. In these cases, f grows polynomially with respect to $\|U\|$ and $e(t) = 0$ exactly. We define the Normalized Extreme Treatment Effect (NETE) as

$$\theta^{\text{NETE}} = \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{Y(1) - Y(0)}{t^\alpha} \mid \|U\| > t \right], \quad (3.2)$$

where α is a known index in Assumption 3.2 from prior knowledge. Note that the previous definition (3.1) is a special case of (3.2) when $\alpha = 0$. The intuition for the scaling factor t^α is that under Assumption 3.2, $\mathbb{E}[Y(d)]$ is of the order $O(\|U\|^\alpha)$ and (3.2) is of the order $O(\mathbb{E}[(\|U\|/t)^\alpha \mid \|U\| > t])$, which is finite if $\alpha < \beta$. (3.2) implies that for a large threshold t , we have $\mathbb{E}[Y(1) - Y(0)] \approx t^\alpha \theta^{\text{NETE}}$. Therefore, θ^{NETE} measures the influence of treatment on the susceptibility of outcome with respect to extreme noise U .

We want to remark that NETE naturally sits at the nexus of two well-studied strands of work, tail-conditional expectations in EVT, and average effects or distributional shifts at extreme quantiles, e.g., ATE and QTE. NETE can be understood as a causal analogue of EVT quantity $\mathbb{E}[Z/t \mid Z > t]$, where Z is a regularly varying variable. It generalizes ATE to the setting of extreme events and aligns with the growth rate given by EVT.

3.2 Extreme Effect Identification and Estimation

The estimand (3.2) is designed to measure the treatment effect under extreme events, i.e., extremely large $\|U\|$. In practice, there may only be a small fraction of extreme samples in the dataset, which creates difficulties for statistical inference. To efficiently estimate the NETE, we leverage the asymptotic independence property of regularly varying variables (See Definition 2.3) to derive a novel identification formula. In particular, we have the following decomposition.

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{Y(1) - Y(0)}{t^\alpha} \mid \|U\| > t \right] &= \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{t^\alpha} \mid \|U\| > t \right] \quad (3.3) \\ &= \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{\|U\|^\alpha} \cdot \left(\frac{\|U\|}{t} \right)^\alpha \mid \|U\| > t \right] \\ &= \lim_{t \rightarrow \infty} \mathbb{E} \left[g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|) \cdot \left(\frac{\|U\|}{t} \right)^\alpha \mid \|U\| > t \right] \quad (3.4) \end{aligned}$$

where we use Assumption 3.2 in the third equality. We can prove that the above quantity equals to

$$\lim_{t \rightarrow \infty} \mathbb{E}[g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|) \mid \|U\| > t] \cdot \lim_{t \rightarrow \infty} \mathbb{E}[\|U\|^\alpha/t^\alpha \mid \|U\| > t].$$

The first factor measures the average effect of treatment across different directions, while the second factor only depends on the norm of the extreme noise, which can be estimated via standard techniques in extreme value theory. We summarize the identification formula in the following proposition.

Proposition 3.3 (Identification). Suppose that U is multivariate regularly varying and Assumption 2.1, 2.2, 3.1, 3.2 hold, we have

$$\theta^{\text{NETE}} = \lim_{t \rightarrow \infty} \mathbb{E}[g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|) \mid \|U\| > t] \cdot \lim_{t \rightarrow \infty} \mathbb{E}[\|U\|^\alpha/t^\alpha \mid \|U\| > t].$$

Proposition 3.3 separates the estimation of NETE into two parts, the expectation of the spectral measure and the index estimation, which facilitates the estimation. While in theory naive identification (3.3) works as well, we found that in practice (3.3) performs poorly (See Section 4 for empirical experiments). One reason is that without properly scaling, the (3.3) suffers from exploding $\|U\|$, causing larger estimation errors.

Inspired by this decomposition, we estimate the two factors separately. To make our framework more flexible, we allow an approximate scaling exponential $\hat{\alpha}_n$ as input in Algorithm 1. $\hat{\alpha}_n$ can be obtained from some prior knowledge or via other heuristics. For the first factor, we design two estimators, the Inverse Propensity Weighting (IPW) and the Doubly Robust (DR) estimators. To derive the estimators, we first randomly split the data into equal halves and use the first half for nuisance estimation, i.e., propensity and outcome. Note that the domain of the propensity function is unbounded due to the presence of U . Generally, it is difficult to estimate such a function given scarce data in the tail. To bypass this barrier, we make the following independence assumption.

Assumption 3.4. (Independence) The extreme variable U is independent of D, X , i.e., $U \perp X, D$.

Under Assumption 3.4, only X affects treatment assignment. We use the first half of data to regress $(X, D, U/\|U\|)$ on $Y/\|U\|^{\hat{\alpha}_n}$ to get (normalized) pseudo-outcome $\hat{g}$ and regress X on D to get an estimation of the propensity function $\hat{p}$. Then, we use the second half for estimation. The IPW and DR estimators are defined in (3.5) and (3.6), respectively.

Notice that the second factor is the α -moment of the random variable $\|U\|/t \mid \|U\| > t$, which converges weakly to a Pareto distribution as t increases to infinity. Therefore, this quantity equals to the α moment of a standard Pareto $1/(1 - \alpha\gamma)$ and the problem can be reduced to estimating the EVI of an asymptotic Pareto distribution. Here, we use the adaptive Hill estimator in (3.7) from

Algorithm 1 Algorithm for NETE Estimation

Require: Dataset $\mathcal{D} = \{(X_i, D_i, Y_i, U_i)\}_{i=1}^n$, threshold t , exponent estimation $\hat{\alpha}_n$, estimator

1: Randomly split $\mathcal{D}$ into two equal parts $\mathcal{D}_1$ and $\mathcal{D}_2$

2: Using $\mathcal{D}_1$, estimate:

a. Propensity function $\hat{p}(x)$ via regression of D on X and clip the output of $\hat{p}(x)$ to the interval $[c, 1 - c]$.

b. Pseudo-outcome regression $\hat{g}(x, d, s)$ by regressing $Y/\|U\|^{\hat{\alpha}_n}$ on $(X, D, U/\|U\|)$

3: Define index set $\mathcal{I} = \{i : \|U_i\| > t, (X_i, D_i, Y_i, U_i) \in \mathcal{D}_2\}$ and set $S_i = U_i/\|U_i\|$ for $i \in \mathcal{I}$

4: **if** estimator = IPW **then**

5: Compute

$$\hat{\eta}_{n,t}^{\text{IPW}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} S_i \left(\frac{D_i}{\hat{p}(X_i)} - \frac{1 - D_i}{1 - \hat{p}(X_i)} \right). \quad (3.5)$$

6: **else if** estimator = DR **then**

7: Compute

$$\hat{\eta}_{n,t}^{\text{DR}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\hat{g}(X_i, 1, S_i) - \hat{g}(X_i, 0, S_i) + \frac{D_i - \hat{p}(X_i)}{\hat{p}(X_i)(1 - \hat{p}(X_i))} (Y_i/\|U_i\|^{\hat{\alpha}_n} - \hat{g}(X_i, D_i, S_i)) \right]. \quad (3.6)$$

8: **end if**

9: Compute adaptive Hill estimator on $\{\|U_i\| : i \in \mathcal{I}\}$:

$$\hat{\gamma}_n = \hat{\gamma}(k) = \frac{1}{k} \sum_{j=1}^k \log \frac{\|U_{(j)}\|}{\|U_{(k+1)}\|}, \quad \hat{\mu}_n = \frac{1}{1 - \hat{\alpha}_n \hat{\gamma}_n}, \quad (3.7)$$

where $\|U_{(1)}\| \geq \dots \geq \|U_{(k+1)}\|$ and k is chosen by

$$k = \max \left\{ k \in \{l_n, \dots, n\} \text{ and } \forall i \in \{l_n, \dots, n\}, |\hat{\gamma}(i) - \hat{\gamma}(k)| \leq \frac{\hat{\gamma}(i) r_n(\delta)}{\sqrt{i}} \right\},$$

10: **return** $\hat{\theta}_{n,t}^{\text{estimator}} = \hat{\eta}_{n,t}^{\text{estimator}} \cdot \hat{\mu}_n$.

Boucheron and Thomas [2015], which provides a data-driven method for choosing the threshold. Putting the two estimations together, we get our estimator of the NETE $\hat{\theta}_{n,t} = \hat{\eta}_{n,t} \cdot \hat{\mu}_n$, where the superscript $\cdot$ can be DR or IPW. We summarize our estimators in Algorithm 1.

Remark 3.5. Assumption 3.4 may not always hold in practice. The identification formula (??) still holds regardless of Assumption 3.4. When the assumption is violated, the propensity score based method cannot be applied here. In this case, better estimators may be a simple plug-in estimator or the DR estimator since they do not require accurate propensity estimation.

Remark 3.6. The identification formula Proposition 3.3 and our estimation methods can be easily extended for CATE estimation. We will discuss the extension in the appendix.

3.3 Non-asymptotic Analysis

Up to now, we have worked under very mild regular variation and asymptotic homogeneity conditions, which suffice to prove the consistency of our two-step estimator in the limit $n, t \rightarrow \infty$. However, to obtain non-non-asymptotic, finite-sample deviation bounds for both the spectral-measure term and the tail-index term, we must invoke a more structured tail model. In particular, existing results such as those in Zhang et al. [2023] rely on the fact that, beyond regular variation, the noise vector behaves exactly like a (possibly linearly transformed) Pareto distribution. Although this is admittedly stronger than mere second-order regular variation, it is at present the only framework in which we can directly apply sharp concentration inequalities and Wasserstein-distance bounds for spectral-measure estimation. We therefore make the following Pareto-type assumption.

Assumption 3.7. We assume that the distribution of U comes from the following class of models

$$M = \cup_{k=1}^{\infty} M_k,$$

where $M_k = \{\mathcal{L}(X) : U = AZ, \text{ for } A \in \mathcal{A} \text{ and } \mathcal{L}(Z) \in \widetilde{M}_k\}$. The set of possible distributions for the components Z is

$$\widetilde{M}_k = \left\{ \begin{array}{l} Z \text{ admits a (Lebesgue) density } h(z) \text{ in } \mathbb{R}_+^{d_z} \\ \mathcal{L}(Z) : \left| \frac{h(z) - \beta^m \prod_{i=1}^m (1+z_i)^{-(\beta+1)}}{\beta^m \prod_{i=1}^m (1+z_i)^{-(\beta+1)}} \right| \leq \xi k^{-s}, \forall z \\ h(z) \propto \prod_{i=1}^m (1+z_i)^{-(\beta+1)} \text{ if } \|z\|_1 > \zeta k^{\frac{1-2s}{\beta}} \end{array} \right\}$$

and the set of possible matrices $\mathcal{A}$ is

$$\mathcal{A} = \left\{ A \in \mathbb{R}_+^{d_u \times d_z} : l \leq \min_i \|A_{\cdot i}\|_1 \leq \max_i \|A_{\cdot i}\|_1 \leq u \text{ and } JA \geq \sigma \right\}.$$

Throughout, we assume the constants satisfy $m \geq d \geq 2, 0 < l < 1 < u, 0 < s < 1/2, \sigma > 0, 0 < \xi < 1$, and $\zeta > 0$.

This assumption states that the extreme variable is a linear transformation of an approximate Pareto distribution. The parameter s measures how close Z is to a standard multivariate Pareto distribution. A small s means the distribution is far from Pareto.

With these assumptions, we are ready to state our main theorem, which gives a non-asymptotic rate to our estimand.

Theorem 3.8. Suppose that Assumption 2.1, 2.2, 3.1, 3.2, 3.4, 3.7 hold, $\alpha < \beta$, where α and β are defined in Assumption 3.2 and Assumption 3.7 respectively. Furthermore, for any fixed t , with probability at least $1 - \delta$,

$$\begin{aligned} |p(X) - \widehat{p}(X)| &\leq R_p(n, \delta), |\widehat{\alpha}_n - \alpha| \leq R_\alpha(n, \delta), \\ |\mathbb{E}[Y/\|U\|^\alpha \mid X, D, U/\|U\|, \|U\| > t] - \widehat{g}(X, D, U/\|U\|)| &\leq R_g(n_t, \delta), \end{aligned}$$

where $n_t = \sum_{i=1}^{\lfloor n/2 \rfloor} I(\|U_i\| > t)$ and R_p, R_g, R_α are estimation errors that are monotonically decreasing with respect to sample size. Then, if $n \geq \Theta(\log(1/\delta)t^\beta)$, with probability at least $1 - \delta, \delta \in (0, 1/2)$, we have

$$\begin{aligned} |\widehat{\theta}_{n,t}^{\text{DR}} - \theta^{\text{NETE}}| &\leq O(\sqrt{R_p(n/2, \delta)R_g(nt^{-\beta}, \delta)} + t^{\beta/2}n^{-1/2} + \log(1/\delta)n^{-1/(2+\beta)} \\ &\quad + t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + \log(t)R_\alpha(n, \delta) + e(t)). \end{aligned} \quad (3.8)$$

and

$$\begin{aligned} |\widehat{\theta}_{n,t}^{\text{IPW}} - \theta^{\text{NETE}}| &\leq O(R_p(n/2, \delta) + t^{\beta/2}n^{-1/2} + \log(1/\delta)n^{-1/(2+\beta)} \\ &\quad + t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + \log(t)R_\alpha(n, \delta) + e(t)). \end{aligned} \quad (3.9)$$

The error bound (3.6) consist of the nuisance error $\sqrt{R_p(n/2, \delta)R_g(nt^{-\beta}, \delta)}$, variance $t^{\beta/2}n^{-1/2}$, EVI estimation error $\log(1/\delta)n^{-1/(2+\beta)}$, α error $R_\alpha(n, \delta)$ and bias terms $t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + e(t)$. Similar pattern holds for (3.9). Given this general result, we choose the threshold t in a data-driven way to obtain a better rate. The idea is to use the estimated index to balance the bias and variance terms in (3.8) and (3.9). The following corollary gives the convergence rate in two different regimes.

Corollary 3.9. Under the assumptions of Theorem 3.8, further suppose that

$$R_p(n, \delta) = \Theta(\log(1/\delta)n^{-1/2}), R_g(n, \delta) = \Theta(\log(1/\delta)n^{-1/2}), R_\alpha(n, \delta) = \Theta(\log(1/\delta)n^{-c_\alpha}),$$

for some $c_\alpha > 0$, the following conclusions hold.

1. If $s \in (0, 1/(2 + \max\{1, \beta\}))$, takes $t_n = \Theta(n^{(1-2s)\widehat{\gamma}_n})$, with probability at least $1 - \delta$, we have

$$|\widehat{\theta}_{n,t}^{\text{DR}} - \theta^{\text{NETE}}| = O(e(t_n) + n^{-s} \log(1/\delta) + n^{-c_\alpha} \log(n) \log(1/\delta)).$$

2. If $s \in [1/(2 + \max\{1, \beta\}), 1/2)$, takes $t = \Theta(n^{(\widehat{\gamma}_n/(1+2\min\{1, \widehat{\gamma}_n\}))})$, with probability at least $1 - \delta$, we have

$$|\widehat{\theta}_{n,t}^{\text{DR}} - \theta^{\text{NETE}}| = O(e(t_n) + n^{-1/(2+\max\{\beta, 1\})} \log(1/\delta) + n^{-c_\alpha} \log(n) \log(1/\delta)).$$

Similar results hold for the IPW estimator. Due to limited space, we leave the result for IPW in the appendix. Many common machine learning algorithms, e.g., Lasso, logistic regression, neural networks, can achieve $O(n^{-1/2})$ rate in the assumption of Corollary 3.9. We want to highlight that if $e(t)$ decays fast enough and become negligible compared to the other term and we know the correct scaling exponential α , Corollary 3.9 matches the rate of [Zhang et al., 2023, Theorem 3.1] without prior knowledge on the index β in the Assumption 3.7. Besides, if we have additional prior knowledge on $e(t)$ and c_α , we can adjust the choice of threshold t to achieve a better rate.

Remark 3.10. When the extreme noise is 1-dimensional, the spectral measure is trivially $\delta_{\{1\}}$ and there is no need to estimate the spectral measure. Following a similar argument of Theorem 3.8 and Corollary 3.9, we can obtain a convergence rate of $O(e(t_n) + \log(1/\delta)n^{-1/(2+\beta)} + \log(1/\delta)n^{-c_\alpha})$.

Remark 3.11. Assumption 3.7 may seem restricted at first glance. This assumption is used here because the non-asymptotic result for regularly varying extreme distributions is rare in the literature and the goal of this paper is not to develop a new estimator for the spectral measure. To the best of our knowledge, Zhang et al. [2023] is the only paper that gives such a result under Assumption 3.7. In fact, Assumption 3.7 can be easily replaced by the following two assumptions in our proof. (1) The extreme noise U is regularly varying and its norm $\|U\|$ satisfies the von Mises condition in Boucheron and Thomas [2015]. (2) There exists an upper bound for the bias term $|\mathbb{E}[f(U/\|U\|) \mid \|U\| > t] - \lim_{t \rightarrow \infty} \mathbb{E}[f(U/\|U\|) \mid \|U\| > t]| \leq O(t^{-c_0})$, for some constant $c_0 > 0$ for a fixed Lipschitz function f . We leave this generalization to future work.

4 Experiments

Having established in Section 3 that under our regularity and overlap assumptions the DR- and IPW-based extreme treatment estimators enjoy a provable non-asymptotic error bound, we next evaluate their finite-sample behavior and compare with our estimators with naive estimators that does not consider the regularly varying structure. In what follows, Section 4.1 presents purely synthetic simulations with known NETE. Section 4.2 then moves to a semi-synthetic setting—using real noise from wavesurge datasets—to assess practical performance under realistic complexities.

4.1 Synthetic Dataset

The data generation process we use in this subsection is

$$X \sim \text{Unif}([0, 1]^5), D \sim \text{Ber}(p(X)), \text{ where } p(x) = 1/(1 + \exp(-x^\top b)), \\ Y = \|U\|^\alpha (D + U/\|U\| + \epsilon) + \|U\|^{\alpha/2}, \epsilon \sim \text{Unif}(-1, 1),$$

where $\alpha > 0$ is a constant and $b \sim N(0, 1)$, $A \sim \text{Unif}([1, 2]^{d_u \times d_z})$. We consider two ways of generating the extreme noise. The first one follows Assumption 3.7.

$$Z = (Z_1, \dots, Z_{d_z}), Z_i \sim \text{Pareto}(\beta), U = AZ, A \in \mathbb{R}^{d_u \times d_z}.$$

We also consider a Pareto mixture, i.e., $U = (U_1, \dots, U_{d_u}), U_i \sim 0.5\text{Pareto}(\beta) + 0.5\text{Pareto}(\beta + 1)$. Note that Assumption 3.2 is satisfied with $e(t) = t^{-\alpha/2}$. By Proposition 3.3, we can calculate the ground-truth effect. The graph below shows the Mean Square Error (MSE) with our estimator using different sample sizes. We take different values for α, β in the experiments. In this case, by Proposition 3.3, we know that the ground-truth NETE is $1/(1 - \alpha/\beta)$. We use Mean-Square-Error (MSE), $\mathbb{E}[(\hat{\theta} - \theta^{\text{NETE}})^2]$, to measure the error. As a baseline, we compare our estimator with naive IPW and DR estimators. Naïve-IPW simply applies the standard IPW estimator to the U_i that has norm larger than a threshold t , ignoring any tail-index modeling. Similarly, Naïve-DR augments it with the usual doubly-robust correction term but likewise ignores the Pareto structure. We leave the detailed math formulation of the baseline estimators to the appendix. The thresholds rule in Corollary 3.9 is used in the experiments and we use the same threshold selection rules for all estimators. We estimate the scaling exponential α by doing linear regression $\log(|Y|) \sim \log(\|U\|)$ and use the coefficient of $\log(\|U\|)$ as $\hat{\alpha}_n$. We leave the experiment details to the appendix. Fig. 1 and Fig. 2 show the experiment results. In the following, we use EVT-IPW and EVT-DR to represent $\hat{\theta}_{n,t}^{\text{IPW}}$ and $\hat{\theta}_{n,t}^{\text{DR}}$ in Algorithm 1.

Figure 1 and Fig. 2 show that under different configurations of α, β, d_u, d_z , our estimators generally perform better than the baseline estimators. The reason is that our estimators can make better

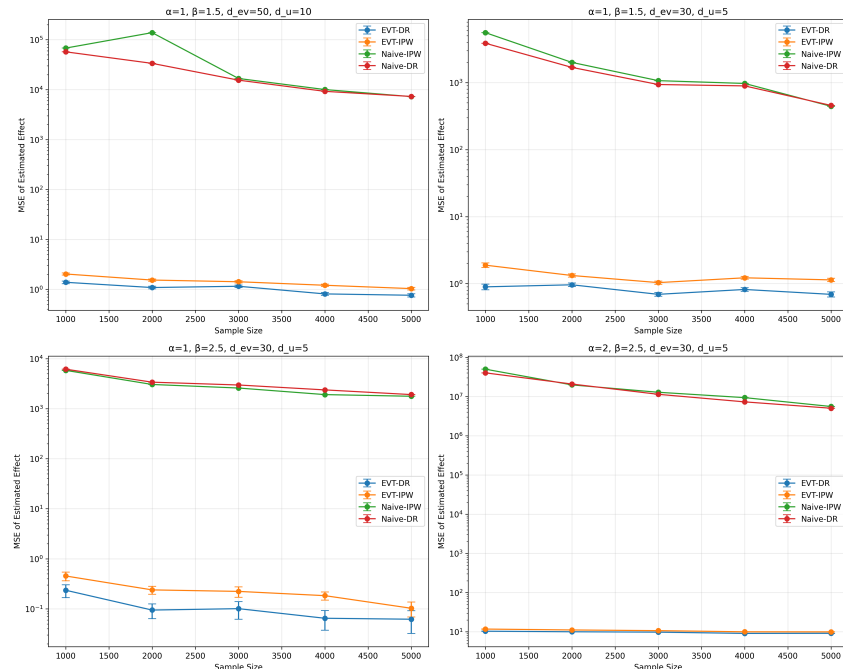


Figure 1: Experiment results of four different configurations when the extreme noise is a linear transformation of Pareto variables. The configurations of upper left, upper right, lower left and lower right are $\alpha, \beta, d_z, d_u = (1, 1.5, 50, 10), (1, 1.5, 30, 5), (1, 2.5, 30, 5)$ and $(2, 2.5, 30, 5)$ respectively. The results are averages of 50 repeated experiments. We use EVT-IPW and EVT-DR to represent $\hat{\theta}_{n,t}^{\text{IPW}}$ and $\hat{\theta}_{n,t}^{\text{DR}}$ in Algorithm 1.

use of the regularly varying structure. In general, EVT-DR achieves the smallest MSE in most experiments and is robust under different configurations. Note that the Pareto mixture does not satisfy Assumption 3.7. Fig. 2 shows that our method still maintain a good performance even if Assumption 3.7 is violated. We also observe that sometimes the MSE increases with more samples in Fig. 2. An explanation for this is that violation of Assumption 3.7 causes the threshold selection rule in Corollary 3.9 not to be applicable and the variance term dominates the error.

4.2 Semi-synthetic Dataset

Now, we use the wavesurge dataset Coles et al. [2001] to create a semi-synthetic dataset for our experiments. The wavesurge dataset has 2894 data points, which contain wave and surge heights at a single location off south-west England. Since wave and surge heights are not in the same scale and may not be positive, we shift the data and normalize each dimension by its 10% quantile. Given the wavesurget dataset, we generate our semi-synthetic dataset in the following way.

$$\begin{aligned} X &\sim \text{Unif}(0, 1), D \sim \text{Ber}(p(X)), \text{ where } p(x) = 1/(1 + \exp(-x^T b)), \\ Y &= (1 - X + D)W^{\alpha_1}S^{\alpha_2} + N(0, 1), \end{aligned} \quad (4.1)$$

where W and S are the height of the wave and surge, respectively. In this experiment, we evaluate how well our proposed EVT-based estimators recover the Normalized Extreme Treatment Effect (NETE) when only limited “short-term” data are available. We split the dataset into a training set (1,000 observations) and a test set (1,894 observations). First, we estimate the NETE on the training set using four estimators. Next, we apply the identification formula from Proposition 3.3 together with (4.1) to obtain a high-fidelity estimate of the NETE on the test set. Because the test-set estimate leverages additional data and the correct tail model, we treat it as a surrogate “ground truth” for comparison. The real-world implications of this experiment is that we can use some short-term data (the training set) to predict long-term and unobserved behavior (the test set).

Table 1 shows the results we get using different estimators. The results show that our EVT-IPW and EVT-DR give estimations that are closer to the test-set estimate than the naive estimators. In particular,

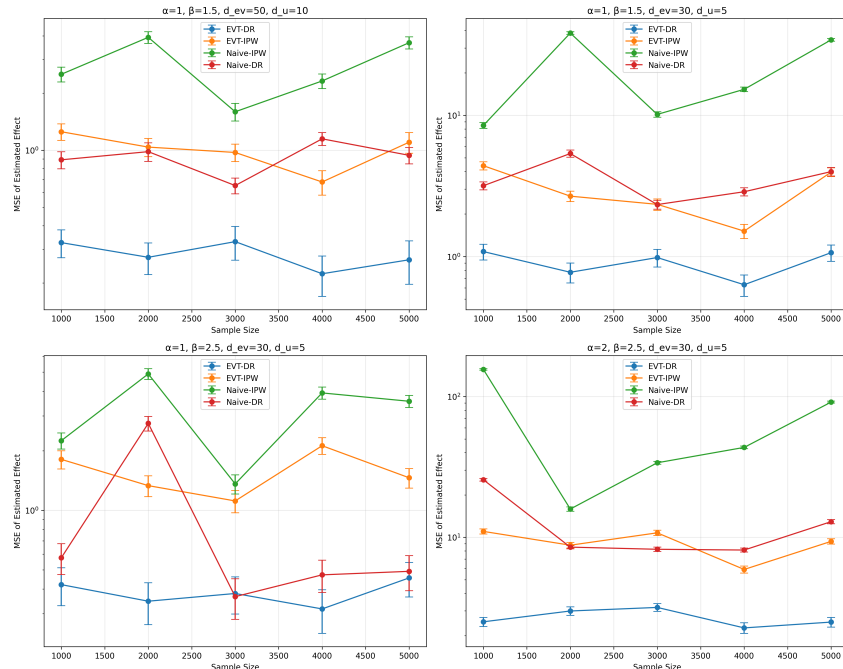


Figure 2: Experiment results of four different configurations when the extreme noise is a Pareto mixture. The configurations of upper left, upper right, lower left and lower right are $\alpha, \beta, d_u = (1, 1.5, 10), (1, 1.5, 5), (1, 2.5, 5)$ and $(2, 2.5, 5)$ respectively. The results are averages of 50 repeated experiments.

the naive estimators consistently overshoot the true NETE by an order of magnitude. In addition, while more extreme tail configurations (e.g. $(1, 3)$) slightly increase variance, the EVT-based methods remain stable, with EVT-DR deviating by at most 0.3 from the test-set estimate. These findings demonstrate that incorporating multivariate extreme value structure via our EVT-IPW and EVT-DR estimators substantially improves finite-sample estimation of treatment effects on rare, tail events, compared both to naive methods.

Table 1: Causal Effect Estimates

(α_1, α_2)	EVT-DR	EVT-IPW	Naive-DR	Naive-IPW	Test-set Estimate
(2, 2)	0.18	0.25	41.93	27.34	0.13
(1, 3)	0.43	0.44	17.04	15.68	0.13
(2.5, 1)	0.13	0.18	31.64	26.06	0.20
(1.5, 1.5)	0.26	0.23	7.91	9.46	0.20

5 Conclusion

In this paper, we tackled the challenge of estimating treatment effects on rare, high-impact events by combining causal inference with extreme value theory. We introduced a new estimand capturing how interventions shift the tail average of outcome distributions and derived a simple identification formula using the spectral-magnitude decomposition of multivariate regular variation. Based on this, we proposed IPW and DR estimators with non-asymptotic error bounds under Pareto-type tails. Simulations and real-data experiments confirmed their advantages over naive estimators when targeting extreme outcomes. This work opens the door to more refined causal analyses in domains such as disaster risk and finance, where tail behavior matters most. A current limitation is our heuristic estimation of the tail index α ; future work will focus on theoretically grounded estimation of α and adaptive threshold selection.

References

- Alberto Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American statistical Association*, 97(457):284–292, 2002.
- Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.
- Ahmed Aloui, Ali Hasan, Yuting Ng, Miroslav Pajic, and Vahid Tarokh. Treatment effects in extreme regimes. *arXiv preprint arXiv:2306.11697*, 2023.
- Marco Avella-Medina, Richard A Davis, and Gennady Samorodnitsky. Kernel pca for multivariate extremes. *arXiv preprint arXiv:2211.13172*, 2022.
- Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.
- Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*, volume 27. Cambridge university press, 1989.
- Juraj Bodik. Extreme treatment effect: Extrapolating causal effects into extreme treatment domain. *arXiv preprint arXiv:2403.11003*, 2024.
- Juraj Bodik, Zbyněk Pawlas, and Milan Paluš. Causality in extremes of time series, August 2023.
- Stéphane Boucheron and Maud Thomas. Tail index estimation, concentration and adaptivity. 2015.
- Victor Chernozhukov and Songzi Du. Extremal Quantiles and Value-at-Risk, May 2006.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. Technical Report 23564, National Bureau of Economic Research, 2017.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, 2018.
- Stephen R Cole and Miguel A Hernán. Fallibility in estimating direct effects. *International journal of epidemiology*, 31(1):163–165, 2002.
- Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.
- Anthony C. Davison and Richard L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 52(3):393–425, 1990.
- David Deuber, Jinzhou Li, Sebastian Engelke, and Marloes H. Maathuis. Estimation and Inference of Extremal Quantile Treatment Effects for Heavy-Tailed Distributions. *Journal of the American Statistical Association*, 119(547):2206–2216, July 2024. ISSN 0162-1459. doi: 10.1080/01621459.2023.2252141.
- Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.
- Eric Gautier and Stefan Hoderlein. A triangular treatment effect model with random coefficients in the selection equation. *arXiv preprint arXiv:1109.0362*, 2011.
- Nadine Gissibl and Claudia Klüppelberg. Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A):2693 – 2720, 2018. doi: 10.3150/17-BEJ941. URL <https://doi.org/10.3150/17-BEJ941>.

- Nicola Gnecco, Nicolai Meinshausen, Jonas Peters, and Sebastian Engelke. Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755–1778, 2021.
- Maike Hohberg, Peter Pütz, and Thomas Kneib. Treatment effects beyond the mean using distributional regression: Methods and guidance. *PloS one*, 15(2):e0226514, 2020.
- Wei Huang, Shuo Li, and Liuhua Peng. Estimation and inference for extreme continuous treatment effects. *Journal of Business & Economic Statistics*, pages 1–26, 2024.
- Joseph D.Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007. doi: 10.1214/07-STS227.
- M. R. Leadbetter. On a basis for ‘Peaks over Threshold’ modeling. *Statistics & Probability Letters*, 12(4):357–362, October 1991. ISSN 0167-7152. doi: 10.1016/0167-7152(91)90107-3.
- Linda Mhalla, Valérie Chavez-Demoulin, and Debbie J Dupuis. Causal mechanism of extreme river discharges in the upper danube basin network. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 69(4):741–764, 2020.
- Quang Huy Nguyen. *Tail distribution of the sums of regularly varying random variables, computations and simulations*. PhD thesis, Lyon 1, 2014.
- James Pickands III. Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131, 1975.
- Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.
- Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.
- Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.
- Richard L. Smith. Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science*, pages 367–377, 1989.
- Fangzhou Su, Wenlong Mou, Peng Ding, and Martin J Wainwright. When is the estimated propensity score better? high-dimensional analysis and bias correction. *arXiv preprint arXiv:2303.17102*, 2023.
- Jérémie Tuganishuri, Chan-Young Yune, Manik Das Adhikari, Seung Woo Lee, Gihong Kim, and Sang-Guk Yum. Prediction of volume of shallow landslides due to rainfall using data-driven models. *Natural Hazards and Earth System Sciences Discussions*, 2024:1–26, 2024.
- Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2(1), 2006. doi: 10.2202/1557-4679.1043.
- Tyler J VanderWeele. Policy-relevant proportions for direct effects. *Epidemiology*, 24(1):175–176, 2013.
- Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- Alice R Zhai and Jonathan H Jiang. Dependence of us hurricane economic loss on maximum wind speed and storm size. *Environmental Research Letters*, 9(6):064019, 2014.
- Xuhui Zhang, Jose Blanchet, Youssef Marzouk, Viet Anh Nguyen, and Sven Wang. Wasserstein-based minimax estimation of dependence in multivariate regularly varying extremes. *arXiv preprint arXiv:2312.09862*, 2023.
- Yichong Zhang. Extremal quantile treatment effects. *The Annals of Statistics*, 46(6B):3707–3740, December 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-AOS1673.

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitation is discussed in the conclusion section in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proofs are given in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: The code will be provided in the camera-ready version.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments can be easily reproduced using a laptop.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This is mainly theory work.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Proofs

A.1 Identification Formula

In this subsection, we derive the identification formula in Proposition 3.3.

Proof. We first prove (3.4). By Assumption 3.2,

$$\begin{aligned}\lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{Y(1) - Y(0)}{t^\alpha} \mid \|U\| > t \right] &= \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{t^\alpha} \mid \|U\| > t \right] \\ &= \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{\|U\|^\alpha} \cdot \left(\frac{\|U\|}{t} \right)^\alpha \mid \|U\| > t \right] \\ &= \lim_{t \rightarrow \infty} \mathbb{E} \left[(g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|) + 2e(t)) \cdot \left(\frac{\|U\|}{t} \right)^\alpha \mid \|U\| > t \right]\end{aligned}$$

We next argue that $\lim_{t \rightarrow \infty} \mathbb{E}[(\|U\|/t)^\alpha \mid \|U\| > t] = \alpha/(\beta - \alpha)$. We have

$$\begin{aligned}\mathbb{E}[(\|U\|/t)^\alpha \mid \|U\| > t] &= 1 + \int_1^\infty P((\|U\|/t)^\alpha \geq r \mid \|U\| > t) \, dr \\ &= 1 + \frac{\int_1^\infty P(\|U\|/t > r^{1/\alpha}) \, dr}{P(\|U\| > t)} \\ &= 1 + \frac{\int_1^\infty \alpha P(\|U\| > rt) r^{\alpha-1} \, dr}{P(\|U\| > t)}\end{aligned}$$

Note that $\|U\|$ is also regularly varying. By Potter's theorem [Bingham et al., 1989, Theorem 1.56], for any $\epsilon > 0$ and sufficiently large t , we have

$$\frac{P(\|U\| > rt)}{P(\|U\| > t)} \leq 2r^{-\beta+\epsilon}.$$

Take $\epsilon > 0$ such that $\alpha - \beta + \epsilon - 1 < -1$, we have for sufficiently large t ,

$$\frac{\int_1^\infty \alpha P(\|U\| > rt) r^{\alpha-1} \, dr}{P(\|U\| > t)} \leq 2 \int_1^\infty \alpha r^{\alpha-\beta+\epsilon-1} \, dr < \infty.$$

Therefore, by the dominance convergence theorem,

$$\mathbb{E}[(\|U\|/t)^\alpha \mid \|U\| > t] \rightarrow 1 + \int_1^\infty \alpha r^{-\beta} \cdot r^{\alpha-1} \, dr = \beta/(\beta - \alpha),$$

which implies

$$\lim_{t \rightarrow \infty} e(t) \mathbb{E}[(\|U\|/t)^\alpha \mid \|U\| > t] = 0$$

and (3.4) holds. We then verify the uniform integrability of function

$$h(U) = \mathbb{E}_X[(g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|))(\|U\|/t)^\alpha].$$

Note that by Assumption 3.2, g is a continuous function on a compact set and thus is bounded by some constant $C > 0$. We have

$$\mathbb{E}[|h(U)| \mid \|U\| > t] \leq 2C \mathbb{E}[(\|U\|/t)^\alpha \mid \|U\| > t].$$

We have proven that the Right Hand Side (RHS) converges to a constant as $t \rightarrow \infty$, which implies that $\mathbb{E}[|h(U)| \mid \|U\| > t]$ is uniformly bounded. We conclude that $h(U)$ is uniformly integrable. By the uniform-integrability convergence theorem,

$$\begin{aligned}\lim_{t \rightarrow \infty} \mathbb{E} \left[(g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|)) \cdot \left(\frac{\|U\|}{t} \right)^\alpha \mid \|U\| > t \right] &= \mathbb{E}_{(r, \theta) \sim \mathcal{L}}[(g(X, 1, \theta) - g(X, 0, \theta))r^\alpha] \\ &= \mathbb{E}_{\theta \sim \mathcal{L}}[(g(X, 1, \theta) - g(X, 0, \theta))] \mathbb{E}_{r \sim \mathcal{L}}[r^\alpha].\end{aligned}$$

where $\mathcal{L}$ is the limiting distribution of $(\|U\|/t, U/\|U\|) \mid \|U\| > t$ and we use the asymptotic independent property of regularly varying distributions (Definition 2.3). $\square$

A.2 Non-asymptotic Analysis

To obtain a convergence rate for the estimator $\hat{\theta}_n^t$, we first analyze the rate of the two factors $\hat{\gamma}_n$ and $\hat{\eta}_{n,t}^{\text{DR}}$.

Lemma A.1. Under Assumption 3.1, 3.7 with probability at least $1 - \delta$, for sufficiently large n , we have

$$|\gamma - \hat{\gamma}_n| \leq O\left(\left(\frac{\log(2/\delta)}{n}\right)^{1/(2+\beta)}\right),$$

where $\gamma = 1/\beta$ is the EVI of U .

Proof. We adopt the non-asymptotic analysis of the adaptive Hill estimator for EVI in [Boucheron and Thomas, 2015]. In the paper, the author adopts an adaptive estimator, choosing k to be

$$k = \max \left\{ k \in \{l_n, \dots, n\} \text{ and } \forall i \in \{l_n, \dots, n\}, |\hat{\gamma}(i) - \hat{\gamma}(k)| \leq \frac{\hat{\gamma}(i)r_n(\delta)}{\sqrt{i}} \right\}$$

where $\hat{\gamma}(i) = \frac{1}{i} \sum_{j=1}^i \log \frac{\|U_{(j)}\|}{\|U_{(i+1)}\|}$ and $r_n(\delta)$ scales like $\sqrt{\log((2/\delta) \log(n))}$. First, we verify the von Mises conditions in Boucheron and Thomas [2015] under Assumption 3.7. Let F be the CDF of $\|U\|$. By Assumption 3.7, we know that

$$g(z) = c\alpha^m \prod_{i=1}^m (1 + z_i)^{-\alpha-1}, \|z\|_1 > \zeta k^{(1-2s)/\alpha}.$$

Note that

$$|c - 1| = \left| \frac{g(z) - \alpha^m \prod_{i=1}^m (1 + z_i)^{-\alpha-1}}{\alpha^m \prod_{i=1}^m (1 + z_i)^{-\alpha-1}} \right| \leq \xi k^{-s} \leq \xi.$$

Let $\tilde{Z}_1, \dots, \tilde{Z}_m \sim \alpha c^{1/m} (1 + z)^{-\alpha-1}$, $z \geq c^{1/(m\alpha)} - 1$. Then, we verify the upper bound for the von Mises function, i.e., $\sup_{s \geq t} |\eta(s)| \leq O(t^\rho)$ for some $\rho < 0$, where η is the von Mises function.

$$\eta(t) = \frac{tU'(t)}{U(t)} - \frac{1}{\beta} \tag{A.1}$$

$$= \frac{1}{tU(t)f(U(t))} - \frac{1}{\beta}, \tag{A.2}$$

where $f(t)$ is the density function of $\sum_{i=1}^m a_i \tilde{z}_i$. By [Nguyen, 2014, Theorem 2.1], we have that when $\|U\|_1 > \max_i \{a_i\} \zeta k^{(1-2s)/\alpha}$,

$$f(t) = C\beta t^{-\beta-1} (1 + D(1 - 1/\beta)t^{-1} + o(t^{-1})). \tag{A.3}$$

Then,

$$\bar{F}(t) = 1 - F(t) = Ct^{-\beta} (1 + Dt^{-1} + o(t^{-1}))$$

and

$$U(t) = C^{1/\beta} t^{1/\beta} (1 + DC^{-1/\beta} t^{-1/\beta} / \beta + o(t^{-1/\beta})). \tag{A.4}$$

Plug in (A.3) and (A.4) into (A.2), we get

$$\eta(t) = \frac{1}{\beta(1 - DC^{-1/\beta} t^{-1/\beta} + o(t^{1/\beta}))} - \frac{1}{\beta} = O(t^{-1/\beta}).$$

Therefore, the growth rate of the von Mises function is bounded. By [Boucheron and Thomas, 2015], with probability at least $1 - \delta$, we have

$$|\gamma - \hat{\gamma}_n| \leq O\left(\left(\frac{\log(2/\delta)}{n}\right)^{1/(2+\beta)}\right),$$

□

Lemma A.2. Under the assumption of Theorem 3.8, with probability at least $1 - \delta$, we have

$$|\hat{\eta}_{n,t}^{\text{DR}} - \eta| \leq O(\sqrt{R_p(n/2, \delta)R_g(t^{-\beta}n, \delta)} + t^{\beta/2}n^{-1/2} + t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + e(t) + \log(t)R_\alpha(n, \delta)).$$

Proof. Let

$$\eta^t = \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{\|U\|^{\hat{\alpha}_n}} \mid \|U\| > t \right],$$

We have the following decomposition

$$|\hat{\eta}_{n,t}^{\text{DR}} - \eta| \leq |\hat{\eta}_{n,t}^{\text{DR}} - \eta^t| + |\eta^t - \eta|.$$

The first term comes from the standard statistical error of DR estimator, while the second term is the bias term caused by the finite threshold. For the first term, by standard DML theory [Foster and Syrgkanis, 2023], we have

$$|\hat{\eta}_{n,t}^{\text{DR}} - \eta^t| \leq O \left(\sqrt{R_p(n/2, \delta)R_g(n_t, \delta)} + n_t^{-1/2} \right),$$

where $n_t = \sum_{i=1}^{n/2} I(\|U_i\| > t)$ is a random variable. By Bernstein's inequality, with probability at least $1 - \delta$,

$$n_t - n\mathbb{P}(\|U\| > t)/2 \geq O \left(\log(1/\delta) + \sqrt{n\mathbb{P}(\|U\| > t) \log(1/\delta)} \right)$$

Therefore, with the same probability, when $n \geq \Theta(\log(1/\delta)t^\beta)$.

$$n_t \geq \frac{1}{4}n\mathbb{P}(\|U\| > t) = \Theta(nt^{-\beta})$$

and we have

$$|\hat{\eta}_{n,t}^{\text{DR}} - \eta^t| \leq O \left(\sqrt{R_p(n/2, \delta)R_g(nt^{-\beta}, \delta)} + t^{\beta/2}n^{-1/2} \right), \quad (\text{A.5})$$

where we use the monotonicity of R_p, R_g .

For the second term (the bias term),

$$\begin{aligned} |\eta^t - \eta| &= \left| \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{\|U\|^{\hat{\alpha}_n}} \mid \|U\| > t \right] - \mathbb{E}[g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|)] \right| \\ &\leq \left| \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{\|U\|^\alpha} - g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|) \mid \|U\| > t \right] \right| \\ &\quad + |\mathbb{E}[g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|)] - \mathbb{E}[g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|) \mid \|U\| > t]| \\ &\quad + \left| \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{\|U\|^\alpha} (1 - \|U\|^{\alpha - \hat{\alpha}_n}) \right] \right| \\ &\leq 2e(t) + |\mathbb{E}[g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|)] - \mathbb{E}[g(X, 1, U/\|U\|) - g(X, 0, U/\|U\|) \mid \|U\| > t]| \\ &\quad + \left| \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{\|U\|^\alpha} (1 - \|U\|^{\alpha - \hat{\alpha}_n}) \mid \|U\| > t \right] \right| \end{aligned}$$

where we use Assumption 3.2 in the last equality. By the error rate assumption in Theorem 3.8,

$$\begin{aligned} \left| \mathbb{E} \left[\frac{f(X, 1, U) - f(X, 0, U)}{\|U\|^\alpha} (1 - \|U\|^{\alpha - \hat{\alpha}_n}) \mid \|U\| > t \right] \right| &\leq C\mathbb{E} \left[|1 - \|U\|^{\alpha - \hat{\alpha}_n}| \mid \|U\| > t \right] \\ &= O(\log(t)R_\alpha(n, \delta)). \end{aligned}$$

Since g is L -Lipschitz continuous, the second term is upper bounded by Wasserstein distance $LW_1(\mathcal{L}_{U/\|U\|}^t, \mathcal{L}_{U/\|U\|})$, where $\mathcal{L}_{U/\|U\|}^t$ is the distribution of $U/\|U\|$ conditioning on $\|U\| > t$ and $\mathcal{L}_{U/\|U\|}$ is its limiting distribution as $t \rightarrow \infty$. Therefore, we have

$$\begin{aligned} |\eta^t - \eta| &\leq 2e(t) + LW_1(\mathcal{L}_{U/\|U\|}^t, \mathcal{L}_{U/\|U\|}) + O(\log(t)R_\alpha(n, \delta)) \\ &\leq 2e(t) + O(t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + \log(t)R_\alpha(n, \delta)), \end{aligned} \quad (\text{A.6})$$

where we use [Zhang et al., 2023, Proposition 3.1] in the last inequality to upper bound the bias term $W_1(\mathcal{L}_{U/\|U\|}^t, \mathcal{L}_{U/\|U\|})$. Combing (A.5) and (A.6), we get

$$|\hat{\eta}_n^t - \eta| \leq O(\sqrt{R_p(n/2, \delta)R_g(t^{-\beta}n, \delta)} + t^{\beta/2}n^{-1/2} + t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + e(t) + \log(t)R_\alpha(n, \delta)).$$

□

Lemma A.3. Under the assumption of Theorem 3.8, with probability at least $1 - \delta$, we have

$$|\hat{\eta}_{n,t}^{\text{IPW}} - \eta| \leq O(R_p(n/2, \delta) + t^{\beta/2}n^{-1/2} + t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + e(t)).$$

Proof. Similar to Lemma A.2, we have the following decomposition.

$$|\hat{\eta}_{n,t}^{\text{IPW}} - \eta| \leq |\hat{\eta}_{n,t}^{\text{IPW}} - \eta^t| + |\eta^t - \eta|.$$

Term $|\eta^t - \eta|$ can be bounded in the same way as in the proof of Lemma A.2. By [Su et al., 2023, Theorem 1], we have

$$|\hat{\eta}_{n,t}^{\text{IPW}} - \eta^t| \leq O(R_p(n/2, \delta) + n_t^{-1/2}) = O(R_p(n/2, \delta) + t^{\beta/2}n^{-1/2} + \log(t)R_\alpha(n, \delta)).$$

The rest of the proof is similar. □

Now we are ready to prove Theorem 3.8.

Proof of Theorem 3.8. Note that by the asymptotic independence property of regularly varying distribution,

$$\begin{aligned} |\hat{\theta}_{n,t}^{\text{DR}} - \theta^{\text{NETE}}| &= |\hat{\eta}_{n,t}^{\text{DR}} \cdot \hat{\mu}_n - \eta \cdot \mu| \\ &\leq |\mu| \cdot |\hat{\eta}_{n,t}^{\text{DR}} - \eta| + |\hat{\eta}_{n,t}^{\text{DR}}| \cdot |\hat{\mu}_n - \mu|. \end{aligned} \quad (\text{A.7})$$

By Lemma A.1 and A.2, with high probability, $\hat{\gamma}_n$ and $|\hat{\eta}_{n,t}^{\text{DR}}|$ is bounded. Note that by Lemma A.1,

$$|\hat{\mu}_n - \mu| = \left| \frac{1}{1 - \hat{\alpha}_n \hat{\gamma}_n} - \frac{1}{1 - \alpha \gamma} \right| = O(|\hat{\alpha}_n - \alpha| + |\hat{\gamma}_n - \gamma|) = O(\log(1/\delta)n^{-1/(2+\beta)} + R_\alpha(n, \delta))$$

Therefore, by Lemma A.2 and (A.7),

$$\begin{aligned} |\hat{\theta}_{n,t}^{\text{DR}} - \theta^{\text{NETE}}| &\leq O(\sqrt{R_p(n/2, \delta)R_g(nt^{-\beta}, \delta)} + t^{\beta/2}n^{-1/2} + \log(1/\delta)n^{-1/(2+\beta)} \\ &\quad + t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + \log(t)R_\alpha(n, \delta) + e(t)). \end{aligned}$$

The bound for $\hat{\theta}_{n,t}^{\text{IPW}}$ can be proven similarly. □

Corollary A.4 (Convergence rate for IPW). Under the assumptions of Theorem 3.8, further suppose that

$$R_p(n, \delta) = \Theta(\log(1/\delta)n^{-1/2}), R_g(n, \delta) = \Theta(\log(1/\delta)n^{-1/2}), R_\alpha(n, \delta) = \Theta(\log(1/\delta)n^{-c_\alpha}),$$

for some $c_\alpha > 0$, the following conclusions hold.

1. If $s \in (0, 1/(2 + \max\{1, \beta\}))$, takes $t_n = \Theta(n^{(1-2s)\hat{\gamma}_n})$, with probability at least $1 - \delta$, we have

$$|\hat{\theta}_{n,t}^{\text{IPW}} - \theta^{\text{NETE}}| = O(e(t_n) + n^{-s} \log(1/\delta) + n^{-c_\alpha} \log(n) \log(1/\delta)).$$

2. If $s \in [1/(2 + \max\{1, \beta\}), 1/2)$, takes $t = \Theta(n^{\widehat{\gamma}_n/(1+2\min\{1, \widehat{\gamma}_n\})})$, with probability at least $1 - \delta$, we have

$$|\widehat{\theta}_{n,t}^{\text{IPW}} - \theta^{NE\text{TE}}| = O(e(t_n) + n^{-1/(2+\max\{\beta, 1\})} \log(1/\delta) + n^{-c_\alpha} \log(n) \log(1/\delta)).$$

Proof of Corollary 3.9. By Theorem 3.8 and the error rate assumption in Corollary 3.9, we have

$$\begin{aligned} |\widehat{\theta}_{n,t}^{\text{DR}} - \theta^{NE\text{TE}}| &\leq O(\log(1/\delta)t^{\beta/4}n^{-1/2} + t^{\beta/2}n^{-1/2} + \log(1/\delta)n^{-1/(2+\beta)} \\ &\quad + t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + \log(t) \log(1/\delta)n^{-c_\alpha} + e(t)) \\ &\leq O(\log(1/\delta)t^{\beta/2}n^{-1/2} + \log(1/\delta)n^{-1/(2+\beta)} \\ &\quad + t^{-\min\{1, \beta\}} + t^{-\beta s/(1-2s)} + \log(t) \log(1/\delta)n^{-c_\alpha} + e(t)). \end{aligned}$$

If $s \in (0, 1/(2 + \max\{1, \beta\}))$, we have

$$\begin{aligned} |\widehat{\theta}_{n,t}^{\text{DR}} - \theta^{NE\text{TE}}| &\leq O(\log(1/\delta)t^{\beta/2}n^{-1/2} + \log(1/\delta)n^{-1/(2+\beta)} \\ &\quad + t^{-\beta s/(1-2s)} + \log(t) \log(1/\delta)n^{-c_\alpha} + e(t)). \end{aligned}$$

Takes $t_n = \Theta(n^{(1-2s)\widehat{\gamma}_n})$, we get

$$\begin{aligned} |\widehat{\theta}_{n,t_n}^{\text{DR}} - \theta^{NE\text{TE}}| &\leq O(\log(1/\delta)n^{(1-2s)\widehat{\gamma}_n\beta/2-1/2} + \log(1/\delta)n^{-1/(2+\beta)} \\ &\quad + n^{-\widehat{\gamma}_n\beta s} + \log(t) \log(1/\delta)n^{-c_\alpha} + e(t_n)). \end{aligned} \quad (\text{A.8})$$

By Lemma A.1, we have

$$|\widehat{\gamma}_n - \gamma| = |\widehat{\gamma}_n - 1/\beta| \leq O(\log(1/\delta)n^{-1/(2+\beta)}).$$

Therefore, $n^{\widehat{\gamma}_n\beta} = 1 + O(n^{-1/(2+\beta)})$. Plug this bound into (A.8) and we can get the results. Similarly, if $s \in [1/(2 + \max\{1, \beta\}), 1/2)$, we have

$$\begin{aligned} |\widehat{\theta}_{n,t}^{\text{DR}} - \theta^{NE\text{TE}}| &\leq O(\log(1/\delta)t^{\beta/2}n^{-1/2} + \log(1/\delta)n^{-1/(2+\beta)} \\ &\quad + t^{-\min\{1, \beta\}} + \log(t) \log(1/\delta)n^{-c_\alpha} + e(t)). \end{aligned}$$

Take $t_n = \Theta(\widehat{\gamma}_n/(1 + 2\min\{1, \widehat{\gamma}_n\}))$, we get the results. $\square$

B Experiment

In this section, we introduce some details of our experiments and provide additional experiments regarding the sensitivity of the algorithm with respect to the scaling parameter α .

B.1 Implementation Details

The two baseline estimators we consider are naive-IPW:

$$\widehat{\theta}_{n,t}^{\text{Naive-IPW}} = \frac{1}{t^\alpha n_t} \sum_{i > n/2: \|U_i\| \geq t} Y_i \left(\frac{D_i}{\widehat{p}(X_i)} - \frac{1 - D_i}{1 - \widehat{p}(X_i)} \right).$$

and naive-DR:

$$\widehat{\eta}_{n,t}^{\text{Naive-DR}} = \frac{1}{t^\alpha n_t} \sum_{i > n/2: \|U_i\| \geq t} \left[\widehat{g}(X_i, 1, U_i) - \widehat{g}(X_i, 0, U_i) + \frac{D_i - \widehat{p}(X_i)}{\widehat{p}(X_i)(1 - \widehat{p}(X_i))} (Y_i - \widehat{g}(X_i, D_i, U_i)) \right],$$

where $n_t = \sum_{i=\lfloor n/2 \rfloor + 1}^n I(\|U_i\| \geq t)$ and $\widehat{p}(X)$ and $\widehat{g}(\cdot)$ are the estimated propensity function and the outcome function respectively. The nuisance estimation of $\widehat{g}$ is obtained by running a regression $Y \sim (X, D, U)$. We clip the propensity to $[10^{-4}, 1 - 10^{-4}]$ to ensure the overlap assumption (Assumption 2.2). $\epsilon \sim \text{Unif}([-1, 1])$ in the data generation in synthetic experiments. We use sample splitting in our experiment, using the first half for nuisance estimation. In the experiment, we use the same threshold t for all estimators, which is given by Corollary 3.9. To choose the threshold, we

first use the adaptive Hill estimator Boucheron and Thomas [2015] to get an estimation of EVI $\hat{\gamma}_n$ and then set the threshold to be $t = 0.25n^{(\hat{\gamma}_n/(1+2\min\{1,\hat{\gamma}_n\}))}$ as in Theorem 3.8. The approximate exponential $\hat{\alpha}_n$ is coefficient of $\log(\|U\|)$ in linear regression $\log(|Y|) \sim \log(\|U\|)$. For the adaptive Hill estimator Boucheron and Thomas [2015], we follow authors' choice for hyperparameters and choose $l_n = 30, r(\delta) = \sqrt{\log \log(n)}$ and

$$k = \min \left\{ k \in \{l_n, \dots, n\} \text{ and } \exists i \in \{l_n, \dots, n\}, |\hat{\gamma}(i) - \hat{\gamma}(k)| > \frac{\hat{\gamma}(i)r_n(\delta)}{\sqrt{i}} \right\} - 1,$$

where $\hat{\gamma}(i) = \frac{1}{i} \sum_{j=1}^i \log \frac{\|U_{(j)}\|}{\|U_{(i+1)}\|}$.

We run logistic regression to estimate the propensity function and use random forest to model the outcome.

For the semi-synthetic experiment, we apply the same hyperparameter as above to estimate NETE. We shift the data to make it positive and normalize each dimension by its 10 % quantile. The Fig. 3 shows the rough distribution of the wavesurge data after these transformations.

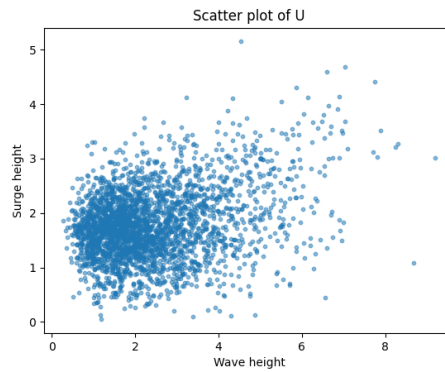


Figure 3: The scatter plot of wavesurge data.

We now describe how we calculate test-set estimation in our experiments. By the data generation process,

$$\begin{aligned} \theta^{\text{NETE}} &= \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{W^{\alpha_1} S^{\alpha_2}}{t^{\alpha_1 + \alpha_2}} \mid \|U\| > t \right] \\ &= \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{W^{\alpha_1} S^{\alpha_2}}{\|U\|^{\alpha_1 + \alpha_2}} \mid \|U\| > t \right] \cdot \frac{1}{1 - (\alpha_1 + \alpha_2)\gamma}, \end{aligned}$$

where we use Proposition 3.3 in the second equality. We know the ground-truth α_1, α_2 and we can estimate the EVI γ using the test set. Suppose the estimated EVI is $\hat{\gamma}$, we set the threshold to $t_n = 0.25n^{(\hat{\gamma}/(1+2\min\{1,\hat{\gamma}\}))}$ and get estimation

$$\hat{\theta}_{\text{test}}^{\text{NETE}} = \mathbb{E}_n \left[\frac{W^{\alpha_1} S^{\alpha_2}}{\|U\|^{\alpha_1 + \alpha_2}} \mid \|U\| > t_n \right] \cdot \frac{1}{1 - (\alpha_1 + \alpha_2)\hat{\gamma}}.$$

B.2 Additional Experiments

We conduct additional experiments under the same setup as in Section 4. Fig. 4 and Fig. 5 show the comparison of EVT-IPW and EVT-DR with and without ground-truth α . We can see that in most cases, their performance is similar, indicating that our heuristic method for estimating α is effective. Indeed, the heuristic approach even gives slightly better MSE.

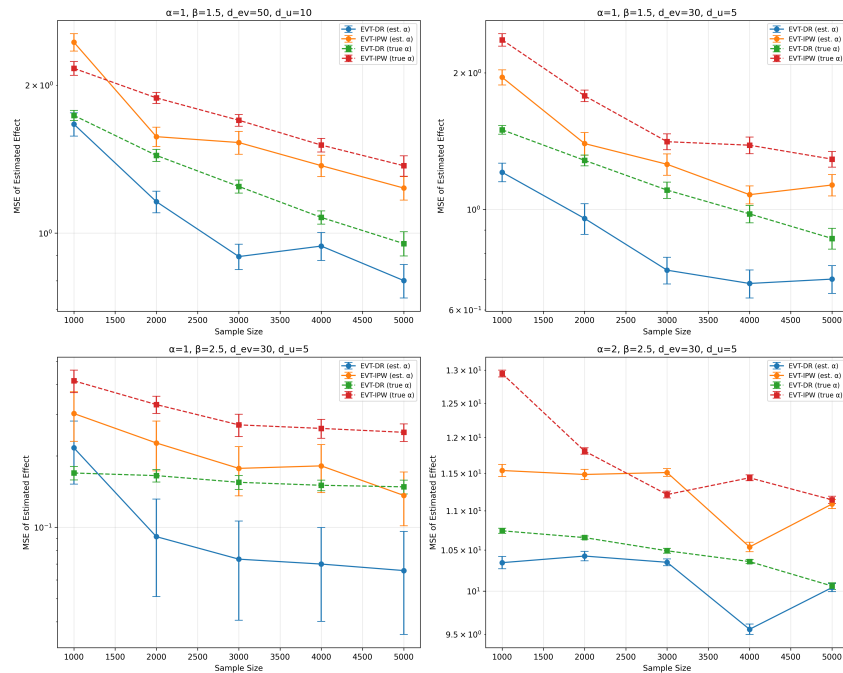


Figure 4: Experiment results of four different configurations when the extreme noise is a linear transformation of Pareto variables. The configurations of upper left, upper right, lower left and lower right are $\alpha, \beta, d_z, d_u = (1, 1.5, 50, 10), (1, 1.5, 30, 5), (1, 2.5, 30, 5)$ and $(2, 2.5, 30, 5)$ respectively. The results are averages of 50 repeated experiments.

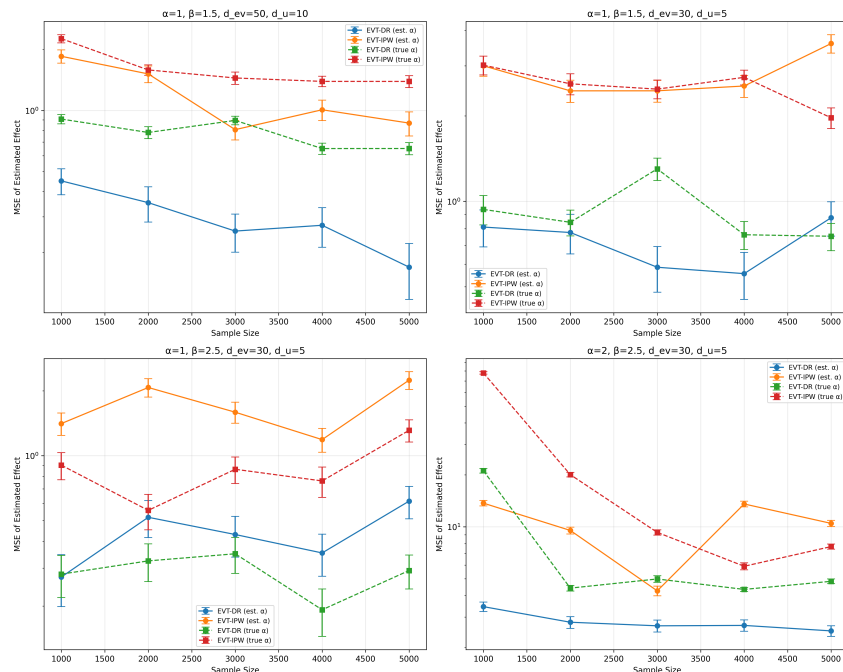


Figure 5: Experiment results of four different configurations when the extreme noise is a Pareto mixture. The configurations of upper left, upper right, lower left and lower right are $\alpha, \beta, d_z, d_u = (1, 1.5, 50, 10), (1, 1.5, 30, 5), (1, 2.5, 30, 5)$ and $(2, 2.5, 30, 5)$ respectively. The results are averages of 50 repeated experiments.

C Extension: Extreme Normalized Heterogeneous Effect Estimation

Proposition C.1 (Identification). Suppose that U is multivariate regularly varying and Assumption 2.1, 2.2, 3.1, 3.2 hold, we have

$$\theta^{\text{CNETE}}(x) = \lim_{t \rightarrow \infty} \mathbb{E}[g(x, 1, U/\|U\|) - g(x, 0, U/\|U\|) \mid \|U\| > t] \cdot \lim_{t \rightarrow \infty} \mathbb{E}[\|U\|^\alpha / t^\alpha \mid \|U\| > t].$$

Proof. The proof is similar to the proof of Proposition 3.3. By Assumption 3.2,

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{Y(1) - Y(0)}{t^\alpha} \mid X = x, \|U\| > t \right] &= \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{f(x, 1, U) - f(x, 0, U)}{t^\alpha} \mid \|U\| > t \right] \\ &= \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{f(x, 1, U) - f(x, 0, U)}{\|U\|^\alpha} \cdot \left(\frac{\|U\|}{t} \right)^\alpha \mid \|U\| > t \right] \\ &= \lim_{t \rightarrow \infty} \mathbb{E} \left[(g(x, 1, U/\|U\|) - g(x, 0, U/\|U\|) + 2e(t)) \cdot \left(\frac{\|U\|}{t} \right)^\alpha \mid \|U\| > t \right]. \end{aligned}$$

By the proof of Proposition 3.3, we know $\lim_{t \rightarrow \infty} \mathbb{E}[(\|U\|/t)^\alpha \mid \|U\| > t] = \alpha/(\beta - \alpha)$. Thus,

$$\lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{Y(1) - Y(0)}{t^\alpha} \mid X = x, \|U\| > t \right] = \lim_{t \rightarrow \infty} \mathbb{E} \left[(g(x, 1, U/\|U\|) - g(x, 0, U/\|U\|)) \cdot \left(\frac{\|U\|}{t} \right)^\alpha \mid \|U\| > t \right].$$

Following the same argument, one can prove the RHS is uniformly bounded. By the dominance convergence theorem,

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{Y(1) - Y(0)}{t^\alpha} \mid X = x, \|U\| > t \right] &= \mathbb{E}_{(r, \theta) \sim \mathcal{L}} [(g(x, 1, \theta) - g(x, 0, \theta))r^\alpha] \\ &= \mathbb{E}_{\theta \sim \mathcal{L}} [g(x, 1, \theta) - g(x, 0, \theta)] \mathbb{E}_{r \sim \mathcal{L}} [r^\alpha]. \end{aligned}$$

□

Similar to NETE estimation, we can separate the estimation of CNETE into two parts. The scaling factor $\mathbb{E}[\|U\|^\alpha / t^\alpha \mid \|U\| > t]$ is the same as the factor in (A.7). For the spectral part, one can use meta learners, e.g, X-learner, S-learner, T-learner, DR-learner, for estimation.