Proceedings of Machine Learning Research vol 247:1–21, 2024 37th Annual Conference on Learning Theory

Limits of Approximating the Median Treatment Effect*

Raghavendra Addanki RADDANKI@ADOBE.COM
Adobe Research

Siddharth Bhandari SIDDHARTH@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Shipra Agrawal and Aaron Roth

Abstract

Average Treatment Effect (ATE) estimation is a well-studied problem in causal inference. However, it does not necessarily capture the heterogeneity in the data, and several approaches have been proposed to tackle the issue, including estimating the Quantile Treatment Effects. In the finite population setting containing n individuals, with treatment and control values denoted by the potential outcome vectors $\mathbf{a}, \mathbf{b}$, much of the prior work focused on estimating median$(\mathbf{a})-$ median$(\mathbf{b})$, as it is easier to estimate than the desired estimand of median$(\mathbf{a}-\mathbf{b})$, called the Median Treatment Effect (MTE). In this work, we argue that MTE is not estimable and detail a novel notion of approximation that relies on the sorted order of the values in $\mathbf{a}-\mathbf{b}$: we approximate the median by a value whose quantiles in $\mathbf{a}-\mathbf{b}$ are close to 0.5 (median). Next, we identify a quantity called *variability* that exactly captures the complexity of MTE estimation. Using this, we establish that when potential outcomes take values in the set $\{0, 1, \ldots, k-1\}$ the worst-case (over inputs $\mathbf{a}, \mathbf{b}$) optimal (over algorithms) approximation factor of the MTE is $\frac{1}{2} \cdot \frac{2k-3}{2k-1}$. Further, by drawing connections to the notions of instance-optimality studied in theoretical computer science, we show that *every* algorithm for estimating the MTE obtains an approximation error that is no better than the error of an algorithm that computes variability, on roughly a per input basis: hence, variability leads to an almost instance optimal approximation algorithm for estimating the MTE. Finally, we provide a simple linear time algorithm for computing the variability exactly. Unlike much prior works, a particular highlight of our work is that we make no assumptions about how the potential outcome vectors are generated or how they are correlated, except that the potential outcome values are k-ary, i.e., take one of k discrete values $\{0, 1, \ldots, k-1\}$.

Keywords: median treatment effect, quantile treatment effect, instance optimality, causal inference

1. Introduction

In the realm of experimentation, much attention is dedicated to constructing efficient estimators for the average treatment effect (ATE), where machine learning based double robust estimation techniques are considered state-of-the-art (Chernozhukov et al., 2018; Kennedy et al., 2017). However, making policy or downstream task decisions only based on ATE, might be sub-optimal, as there is a loss of valuable information from the heterogenous regions in the data. There have been many methods proposed in the literature for tackling this issue (Imai and Ratkovic, 2013; Athey and Imbens, 2016; Ding et al., 2019), and we focus on investigating variations across different outcome distribution quantiles. Unlike ATE, which *averages* out the responses, Quantile Treatment Effects (QTE) (Chernozhukov and Hansen, 2005) provide a more nuanced perspective by unveiling

*The author ordering is alphabetical.

whether effects differ among individuals or outcomes, allowing for a more detailed understanding. This approach proves invaluable in discerning if a treatment is more impactful at specific points in the distribution, facilitating the study of policies affecting diverse population segments, and groups unfairly affected. Several cautionary tales exist in the literature (Heckman et al., 1997), with recent examples related to Netflix's A/B testing oversight of treatment effects disparities among user engagement levels (Bojinov et al., 2020).

We study the finite populating setting on n individuals, where an experimenter assigns the individuals to treatment and control groups. Suppose that the treatment and control values of all the individuals be denoted by $\mathbf{a}$ and $\mathbf{b}$ respectively. Much of the work on QTE estimation has focused on estimating the difference of quantiles of treatment and control (Howard and Ramdas, 2022; Martinez-Taboada and Kennedy, 2023). E.g., for median estimation, the estimand studied is $\mathit{median}(\mathbf{a}) - \mathit{median}(\mathbf{b})$. Although efficient estimators are known, such an estimand has a limitation, as it is not based on the causal effect vector $\mathbf{a} - \mathbf{b}$, which of course is not observable. It might not necessarily capture the causal effect entirely. In contrast, some prior work considered estimands related to $\mathit{median}(\mathbf{a}-\mathbf{b})$, broadly speaking (Heckman et al., 1997; Leqi and Kennedy, 2021; Kallus, 2022; Kennedy, 2023). Recently, Kallus (2022) studied a related problem of identifying the fraction of individuals that are negatively affected (under binary potential outcomes). They provide efficient estimators, but assume the availability of covariates and make distributional assumptions.

In this work, we study the estimation of $\mathit{median}(\mathbf{a}-\mathbf{b})$, henceforth called the Median Treatment Effect (MTE) estimation, without any assumptions, except that the potential outcomes, i.e., the vectors $\mathbf{a}$ and $\mathbf{b}$ are k-ary for some value of k. To the best of our knowledge, this is the first work that tackles the problem in a general setting with minimal or no assumptions. We want to highlight that we *do not* make any distributional assumptions, and the vectors $\mathbf{a}$ and $\mathbf{b}$ can be arbitrarily correlated. Our results contribute to the growing body of work that identifies connections between theoretical computer science and causal inference (Pouget-Abadie et al., 2019; Addanki et al., 2022; Harshaw et al., 2023). In our work, we answer the following questions.

1. **Is Median Estimable?** We show that MTE estimation is infeasible, in stark contrast to ATE estimation. In particular, we show that *any* estimator (algorithm) is unable to distinguish between two distributions from which the treatment and control values are drawn if they have the same marginal distributions. Further, the two distributions can have a large difference in their median values.

2. **Can we approximate MTE?** To overcome the difficulty, we define a notion of approximation that does not rely on the estimate being additively close to the median, but instead approximates the median by being close to in terms of quantiles, i.e., the rank in the sorted order of the values in $\mathbf{a} - \mathbf{b}$. Further, we show that we can approximately estimate MTE under this notion.

3. **Limits of Approximating MTE.** We identify a quantity called *variability*, that essentially captures the complexity of MTE estimation. We show that any algorithm that computes variability obtains the minimum approximation error, and also provide an efficient greedy algorithm for computing the variability.

1.1. Technical Overview

Throughout the overview, we focus on binary outcomes (i.e., $k = 2$), where $\mathbf{a}, \mathbf{b}$ are binary. All our results extend to the situation where we have k-ary outcomes. When $k = 2$, the MTE which is the median of the difference vector $\mathbf{a} - \mathbf{b}$ can be only one of the following values: $\{-1, 0, 1\}$. An experimenter provides us an assignment while ensuring that every individual i is assigned to either the treatment or control group but not both. Hence, we observe either a_i or b_i but not both. Further, the design may be adaptive or non-adaptive, i.e., the experimenter can choose to assign an individual based on all the previously assigned individuals and their responses. We also use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ and for a distribution η we use $\eta^{\otimes n}$ to denote the n-fold product distribution.

1.1.1. IS MEDIAN ESTIMABLE?

Consider two distributions μ_1: $\begin{bmatrix} 1/3 & 0 \\ 1/3 & 1/3 \end{bmatrix}$ and μ_2 : $\begin{bmatrix} 0 & 1/3 \\ 2/3 & 0 \end{bmatrix}$, that encode the joint distributions of the pair $(\mathbf{a}, \mathbf{b})$. So, essentially, we draw n samples from μ_1 or μ_2 to construct $(\mathbf{a}, \mathbf{b})$. Here, the first row encodes the probabilities: For an ith individual drawn from μ_1: $\Pr_{\mu_1}[(a_i, b_i) = (0, 0)] = \mu_1[0, 0] = 1/3$, $\Pr_{\mu_1}[(a_i, b_i) = (0, 1)] = \mu_1[0, 1] = 0$, and so on.

Observe that for both μ_1 and μ_2, the marginal distributions are the same, i.e., $\Pr_{\mu_1}[a_i = 0] = \Pr_{\mu_2}[a_i = 0] = 1/3$ and $\Pr_{\mu_1}[b_i = 0] = \Pr_{\mu_2}[b_i = 0] = 2/3$. Therefore, the ATE of $\mathbf{a} - \mathbf{b}$ for both the distributions is the same and equals $1/3$. However, the MTE for samples drawn from μ_1 is 0, and for μ_2 is 1. This is because the total fraction of $1s$ in a typical sample from μ_2 is $2/3$ and hence contains the median, which is the middle element in the sorted order. However, in μ_1, the fraction of -1s is 0 and combined with the fraction of 0s, which is $\mu_1[0, 0] + \mu_1[1, 1] = 2/3$, it exceeds the $1/2$ threshold, therefore, the MTE for μ_1 is 0. This example illustrates the inherent difficulty in estimating the median from the observed responses in the population. We formalize the notion and provide a brief intuition.

We can think of any algorithm for estimating the MTE as a rooted and labeled decision tree of depth n. Each non-leaf node u, is labeled with a distribution γ_u defined over $\{a, b\} \times [n]$, where a indicates treatment and b indicates control. For each element say (c, i) in the support there are two edges out of u labelled $(c, i, 0)$ and $(c, i, 1)$ respectively. We start from the root, and whenever we are at a non-leaf node u, our algorithm samples (c, i) according to γ_u and assigns the ith individual to treatment if $c = a$ and control if $c = b$. Suppose we see a response value of 0 in individual i, then, we follow the edge $(c, i, 0)$; if ith we observe a response value 1, we follow the edge $(c, i, 1)$. Proceeding in this way we trace a path, say P, from the root to a leaf of the decision tree. The feasibility constraint is that along any path P that has a non-zero probability of being followed, any individual $i \in [n]$ is assigned exactly once. We define such decision trees as *Randomized Feasible Decision Tree* (RFDTs) (see Definitions 14 and 16 for a precise definition). By following a path P, we output an estimate for the MTE say $\widehat{m} \in \{-1, 0, 1\}$ (we can think of the leaf node of P as being labeled with the output). Notice that our definition of RFDTs is general enough to caputer non-adaptive designs. In fact, adaptive designs are those where the distribution γ_u depends only on the depth of u in the decision tree.

A crucial observation is that any RFDT, R_n is incapable of distinguishing between two joint distributions over $\{0, 1\} \times \{0, 1\}$, say η and τ, which have the same marginals, i.e., if we sample $(a, b) \sim \eta$ and call the marginal of a as η_a and of b as η_b (similarly for τ_a and τ_b), then, $\eta_a = \tau_a$ and $\eta_b = \tau_b$. Specifically, any RFDT R_n is unable to distinguish whether the input is drawn from the

distribution $\eta^{\otimes n}$ or $\tau^{\otimes n}$. The reason is that for any path P of R_n the probability that P is followed on the input $(\mathbf{a}, \mathbf{b})$ drawn from $\eta^{\otimes n}$ is a function of the distributions associated with the nodes of P, and the marginal distributions η_a and η_b. Since, this probability only depends on the marginals $\eta_a = \tau_a$ and $\eta_b = \tau_b$, hence the probability of following P on $\tau^{\otimes n}$ remains the same (see Lemma 20 for a detailed proof).

1.1.2. APPROXIMATING THE MEDIAN

Faced with the above handicap, the most refined information an algorithm (RFDT) can obtain from the observed responses is the (approximate) marginals of the control and treatment vectors. Thus any algorithm that estimates the MTE, must in principle use only the marginals. However, the MTEs of treatment and control vectors drawn from $\eta^{\otimes n}$ and $\tau^{\otimes n}$ can be quite different, as shown in the example taken previously. Therefore, we realize that an unbiased estimation of the MTE is impossible. Notice the contrast with ATE estimation, as the ATE value is the same whether the data is generated according to $\eta^{\otimes n}$ and $\tau^{\otimes n}$. This is because the ATE is linear in the outcomes $\mathbf{a}, \mathbf{b}$ but the MTE is not.

To circumvent the above roadblock in estimating the MTE, a natural question is to ask if we can obtain an additive approximation to the MTE, i.e., output an $m \in \mathbb{R}$ such that the MTE is within $\pm\epsilon$ of the MTE for an approximation parameter $\epsilon > 0$. However, for the situation of $k = 2$ (or other values of k) this is not a desirable notion as we already know that the MTE is one of $\{-1, 0, 1\}$, so either we need to set $\epsilon = 1$, which is too large, or know the MTE exactly.

To overcome this, we introduce a notion of approximation that approximates the median by being close to in terms of quantiles, i.e., the rank in the sorted order. More precisely, we define $\widehat{m}$ to be an *ϵ-quantile-approximation* of the median if it lies in the $0.5 \pm \epsilon$-quantile band around the median, denoted by $(\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}$. An ϵ-quantile band around the median is the set of $\epsilon \cdot n$ values of the vector $\mathbf{a} - \mathbf{b}$ in its sorted order, above and below the median.

1.1.3. OPTIMALITY OF MEDIAN APPROXIMATION

To accommodate the above notion of approximation we add a parameter to the output of an algorithm (RFDT) for estimating the MTE. In addition to an estimate $\widehat{m}$ of the MTE, we also output the *width* parameter ϵ, with our guess being that $\widehat{m} \in (\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}$.

How should we measure the performance of any algorithm (RFDT), say R_n (n denotes the number of individuals), that approximately estimates MTE? We measure it via two important parameters: how often is R_n correct, i.e., m is indeed an ϵ-quantile approximation to the MTE, and what is the expected width R_n outputs. More precisely, for some small $\delta > 0$ we desire that for *all inputs* $(\mathbf{a}, \mathbf{b})$ (there is no randomness in the inputs) with probability at least $1 - \delta$ over the internal randomness of R_n we have $\Pr[m \in (\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}]$ where (m, ϵ) denotes the random output of R_n on the input $(\mathbf{a}, \mathbf{b})$. The width of R_n on the input $(\mathbf{a}, \mathbf{b})$, denoted by $\epsilon(\mathbf{a}, \mathbf{b})$ is $\mathbb{E}[\epsilon]$, where the expectation is over the randomness of the algorithm.

Naturally, we desire to construct an RFDT R_n such that it has a small error (δ) and a small width for any input $(\mathbf{a}, \mathbf{b})$. But how small can we make the width? Recall the input distributions μ_1, μ_2 discussed in Section 1.1.1. We know that R_n can't distinguish between $\mu_1^{\otimes n}$ and $\mu_2^{\otimes n}$. Therefore, if upon observing the responses R_n decides to output $\widehat{m} = 0$, then to be correct with high probability it must output the width ϵ as at least $1/6$: this is because for a typical sample $\mathbf{a}, \mathbf{b}$ from $\mu_2^{\otimes n}$ we have that the rank of 0 in the sorted version of $\mathbf{a} - \mathbf{b}$ is $n/3$. Similarly, we

observe that the minimum widths required to be output by R_n for each of the estimates $(-1, 0, 1)$ are $(1/2, 1/6, 1/6)$ respectively. Hence, $(1/6)^{th}$ is in principle a lower bound on the width of R_n for inputs generated from $\mu_1^{\otimes n}$, $\mu_2^{\otimes n}$, and in fact from any $\mu^{\otimes n}$ that shares the same marginals.

We extend and formalize this notion of *minimum median width* of η_a, η_b for any pair of marginal distributions η_a over treatment values and η_b over control values, denoted as $\epsilon^*(\eta_a, \eta_b)$. To compute $\epsilon^*(\eta_a, \eta_b)$ we fix a possible median value $r \in \{-1, 0, 1\}$ and go over all joint distributions η with marginals η_a, η_b, and analyze the maximum deviation of the quantiles that r occupies from $1/2$, i.e., how far is it from the median in the quantile sense, in a typical sample from $\eta^{\otimes n}$. We call this the *width-of-*r for η_a, η_b, denoted as $\epsilon(r, \eta_a, \eta_b)$, and the minimum width-of-r over $r \in \{-1, 0, 1\}$ is $\epsilon^*(\eta_a, \eta_b)$. See Definition 12 for precise definitions. For instance, when η_a was $(1/3, 2/3)$ (denoting the probabilities of being 0 and 1) and η_b was $(2/3, 1/3)$, we saw that the two joint distributions μ_1 and μ_2 (both having the marginals η_a, η_b) already implied that $\epsilon^*(\eta_a, \eta_b) \geq 1/6$: in fact, it will turn out that $\epsilon^*(\eta_a, \eta_b) = 1/6$. For any pair of marginals η_a, η_b, as we have seen previously, any algorithm (RFDT) is unable to distinguish between data generated from $\eta^{\otimes n}$ and $\tau^{\otimes n}$ when η and τ have marginals η_a, η_b: thus, following an argument similar to the previous paragraph, we conclude that in principal $\epsilon^*(\eta_a, \eta_b)$ serves as a lower-bound on the width of any RFDT which is correct with high probability.

To get a handle on the above quantities we define the notion of *variability* (see Definition 11) which is a function of only the marginals and the estimate. Informally, it measures how much the rank of a particular estimate r can vary across joint distributions consistent with the given marginals. For a given $r \in \{-1, 0, 1\}$ and a pair of marginals η_a and η_b the variability of r with respect to η_a, η_b is a 2-tuple of values denoted by $(\nu_l(r, \eta_a, \eta_b), \nu_u(r, \eta_a, \eta_b))$. Let $\mathcal{J}$ be the collection of all joint distributions over $\{0, 1\} \times \{0, 1\}$ which have marginals η_a, η_b. Then, $\nu_l(r, \eta_a, \eta_b)$ is a number in $[0, 1]$ which denotes the supremum over all joint distributions $\eta \in \mathcal{J}$ of the lowest quantile that r occupies in a typical sample of treatment and control values, $\mathbf{a}, \mathbf{b}$, from $\eta^{\otimes n}$, i.e.,

$$\nu_\ell(r, \eta_a, \eta_b) = \sup_{\substack{\eta \in \mathcal{J} \\ (\mathbf{a}, \mathbf{b}) \leftarrow \eta^{\otimes n}}} \frac{|\{i \in [n] \mid (\mathbf{a} - \mathbf{b})_i < r\}|}{n}.$$

Similarly, $\nu_u(r, \eta_a, \eta_b)$ denotes the infimum over all joint distributions $\eta \in \mathcal{J}$ of the highest quantile that r occupies in a typical sample of treatment and control values, $\mathbf{a}, \mathbf{b}$, from $\eta^{\otimes n}$, i.e.,

$$\nu_u(r, \eta_a, \eta_b) = \inf_{\substack{\eta \in \mathcal{J} \\ (\mathbf{a}, \mathbf{b}) \leftarrow \eta^{\otimes n}}} \frac{|\{i \in [n] \mid (\mathbf{a} - \mathbf{b})_i \leq r\}|}{n}.$$

Notice that $\nu_\ell(-1, \eta_a, \eta_b)$ is always 0 and $\nu_u(1, \eta_a, \eta_b)$ is always 1, irrespective of η_a, η_b. To gain some familiarity, below we describe the variability of a particular choice of marginals η_a, η_b. Suppose $\eta_a = (1/3, 2/3)$ and $\eta_b = (2/3, 1/3)$: recall the joint distributions $\mu_1 = \begin{bmatrix} 1/3 & 0 \\ 1/3 & 1/3 \end{bmatrix}$ and $\mu_2 = \begin{bmatrix} 0 & 1/3 \\ 2/3 & 0 \end{bmatrix}$, described in the beginning of Section 1.1.1, consistent with the marginals η_a, η_b. We can argue that $\nu_\ell(1, \eta_a, \eta_b) = 2/3$ and is witnessed by the joint distribution μ_1. Similarly $\nu_\ell(0, \eta_a, \eta_b) = 1/3$ and $\nu_u(0, \eta_a, \eta_b) = 1/3$, both witnessed by μ_2: further, $\nu_u(-1, \eta_a, \eta_b) = 0$ as witnesses by μ_1.

For a pair of marginals η_a, η_b and a median estimate $r \in \{-1, 0, 1\}$, the relationship between $\epsilon(r, \eta_a, \eta_b)$ and variability arises due to the following observation, which in turn also connects $\epsilon^*(\eta_a, \eta_b)$ and variability:

$$\epsilon(r, \eta_a, \eta_b) = \max\{(\nu_l(r, \eta_a, \eta_b) - 1/2)^+, (1/2 - \nu_u(r, \eta_a, \eta_b))^+\}.$$

The observation follows from unpacking the respective concepts, and in fact in the technical sections of the paper we use this observation to define $\epsilon(r, \eta_a, \eta_b)$. However, we illustrate the phenomenon on a specific example: we saw that when $\eta_a = (1/3, 2/3)$ and $\eta_b = (2/3, 1/3)$, $\nu_u(1, \eta_a, \eta_b) = 1$ and $\nu_\ell(1, \eta_a, \eta_b) = 2/3$. Therefore, $\epsilon(1, \eta_a, \eta_b) = 1/6$: which essentially says that in a typical sample $\mathbf{a}, \mathbf{b}$ from any $\eta^{\otimes n}$, where η is a joint distribution with marginals η_a, η_b, the maximum deviation of the quantiles that $r = 1$ occupies from $1/2$ is $1/6$. But we know this to be true because we know that the top $1/3^{rd}$ of the entries of sorted $\mathbf{a} - \mathbf{b}$ have to be 1, when $\mathbf{a}, \mathbf{b}$ is a typical sample from $\eta^{\otimes n}$: hence, 1 always occupies the $2/3^{rd}$-quantile irrespective of η, and hence its quantile deviation from $1/2$ is at most $2/3 - 1/2 = 1/6$.

Using the above connections, we show that the minimum median width is always upper bounded by $\frac{1}{2} \cdot \frac{(2k-3)}{(2k-1)}$ when $\mathbf{a}, \mathbf{b}$ are k-ary (see Lemma 22). For $k = 2$, this value turns out to be $1/6$ and the distributions described above are thus also the hard instances. Hard instances that obtain the upper bound on the minimum width for every k are also detailed in Section 3.

So far, we have shown that variability (or minimum median width) essentially serves as a limit to MTE estimation. We argue, perhaps surprisingly that, variability also lends a tight computational handle on the MTE, even without assumptions on the treatment and control vectors, $\mathbf{a}, \mathbf{b}$, in the following sense. First, in Section 4, we provide an algorithm that computes variability exactly in time $O(k)$. We employ a greedy strategy in trying to identify an appropriate joint distribution that explains the marginals and also allows us to compute the variability. Second, we show the algorithm, called MEDIAN-ESTIMATOR (Algorithm 1), that outputs the minimum median width of the marginals estimated from the observed responses is instance optimal (Theorem 21). We want to emphasize that our algorithm makes no distributional assumption on the inputs.

1.2. Our Contributions

In this section, we highlight our key results. For the sake of brevity, we assume the following notation for all the below-stated results. Let PO_k denote the set of k-ary potential outcomes, η be any joint distribution on $\text{PO}_k \times \text{PO}_k$ with marginals η_a and η_b for the treatment and control respectively, and $\epsilon^*(\eta_a, \eta_b)$ denote the minimum median width (see Definition 12).

In the *instance-optimality* result described below, we show that any algorithm, encoded using an RFDT R_n, has an expected error that is at least the minimum median width $\epsilon^*(\eta_a, \eta_b)$. In other words, $\epsilon^*(\eta_a, \eta_b)$ essentially captures the minimum possible approximation error of any algorithm.

Theorem 1 (Lower Bound on Width of any RFDT (Informal): see Theorem 21) *Let R_n be an RFDT with width $\epsilon^{R_n}(\mathbf{a}, \mathbf{b})$ where $(\mathbf{a}, \mathbf{b})$ are the treatment and control vectors. If for all $(\mathbf{a}, \mathbf{b}) \in \text{PO}_k^n \times \text{PO}_k^n$, with high probability R_n outputs an median estimate $\widehat{m}$ and width ϵ such that $\widehat{m} \in (\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}$, then, we have:*

$$\mathop{\mathbb{E}}_{(\mathbf{a},\mathbf{b}) \leftarrow \eta^{\otimes n}} [\epsilon^{R_n}(\mathbf{a}, \mathbf{b})] \gtrsim \epsilon^*(\eta_a, \eta_b)$$

Let $\psi_k = \frac{1}{2} \cdot \frac{2k-3}{2k-1}$. The theorem below states that the minimum median width $\epsilon^*(\eta_a, \eta_b)$ is always upper bounded by ψ_k and also show a hardness result by identifying marginals $\bar{\eta}_a, \bar{\eta}_b$ that satisfy the upper bound.

Theorem 2 (Tight Bounds for Minimum Median Width (Informal): see Theorem 23) *For any η_a, η_b, we have: $\epsilon^*(\eta_a, \eta_b) \leq \psi_k$. And $\exists \bar{\eta}_a, \bar{\eta}_b$, such that $\epsilon^*(\bar{\eta}_a, \bar{\eta}_b) \geq \psi_k$, which implies $\epsilon^*(\bar{\eta}_a, \bar{\eta}_b) = \psi_k$. Hence, when $k = 2$, by Theorem 1 and the above, $(1/6)^{th}$ is the fundamental limit on approximating the MTE in the quantile sense.*

We want to highlight to the reader that the proof of the upper bound $\epsilon^*(\eta_a, \eta_b) \leq \psi_k$ in Theorem 23 uses a simple and elegant argument based on the expected difference $\mathbb{E}_{X \leftarrow \eta_a}[X] - \mathbb{E}_{Y \leftarrow \eta_b}[Y]$, which by linearity of expectation is $\mathbb{E}_{X,Y \leftarrow \eta}[X - Y]$ across all joint distributions η with marginals η_a, η_b.

Finally, given observed potential outcome vectors (obtained using a Bernoulli design) $\widehat{\mathbf{a}}, \widehat{\mathbf{b}}$, we provide a highly efficient greedy algorithm with linear running time, that returns a median estimate with a width roughly equal to the minimum median width of the empirical frequency distribution. In light of Theorem 1, we know that essentially this is optimal.

Theorem 3 (Median Estimator Algorithm (Informal): see Theorem 24) *Consider the* Bernoulli Design *for assignment, i.e., each individual is assigned to control/treatment groups independently with probability $1/2$. Fix arbitrary treatment and control vectors $\mathbf{a}, \mathbf{b}$ and let the observed response vectors be $\widehat{\mathbf{a}}, \widehat{\mathbf{b}}$. Then, there is an algorithm that returns a median estimator $\widehat{m}$, with width $\epsilon \approx \epsilon^*(\eta_a, \eta_b)$ and runs in $O(n + k^2)$ time. Further, with high probability $\widehat{m} \in (\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}$.*

Note: Our variability estimation method can be extended to an online setting, provided we have accurate approximations for the marginals. Thus our median (or quantile) estimation naturally extends to an online setting, with the vectors $\mathbf{a}, \mathbf{b}$ revealed gradually.

2. Preliminaries and Problem Formulation

We use a (and b) to denote the treatment (and control) value for a particular individual and $\mathbf{a}$ (and $\mathbf{b}$) to denote the treatment (and control) values of all the n individuals. As we do not observe all the treatment and control response values for every individual, we denote the partially observed responses using $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$. Let $[n] = \{1, 2, \ldots, n\}$ and let $\forall x \in \mathbf{R}$, $(x)^+ = \max\{x, 0\}$. For a distribution η we use $\eta^{\otimes n}$ to denote the n-fold product distribution.

Definition 4 (k-ary outcomes) *Let $\mathrm{PO}_k \stackrel{def}{=} \{0, 1, \cdots, k-1\}$ be the set of all k-ary outcomes for both treatment and control groups, i.e., $\mathbf{a}$ and $\mathbf{b}$. Each potential outcome (PO) value for an individual $i \in [n]$ lies in the set PO_k, i.e., $a_i, b_i \in \mathrm{PO}_k \quad \forall i \in [n]$. Therefore, $\mathrm{ITE}_i \in \{-(k-1), \cdots, k-1\}$ $\forall i \in [n]$.*

We make the following assumptions that are common in causality literature.

Assumption 5 (Independence) *The experimental design is independent of the values in $\mathbf{a}, \mathbf{b}$.*

Assumption 6 (SUTVA) *The treatment assignment of a single individual does not affect the outcomes of other units.*

Now, we begin with the definition of a quantile of any vector.

Definition 7 (Quantiles of a vector) *Let $\mathbf{v} \in \mathbb{R}^n$ be a vector, and let $\ell, u \in [0,1]$ be numbers such that $\ell \leq u$. For any $r \in \mathbb{R}$ define $q_\ell(r, \mathbf{v}) \stackrel{def}{=} \frac{|\{i \in [n] | \mathbf{v}_i < r\}|}{n}$ and $q_u(r, \mathbf{v}) \stackrel{def}{=} \frac{|\{i \in [n] | \mathbf{v}_i \leq r\}|}{n}$. Finally, define the set of ℓ-to-u* quantiles of $\mathbf{v}$ *as:*

$$\left(\mathbf{v}\right)_\ell^u \stackrel{def}{=} \{r \in \mathbb{R} \mid [\ell, u] \cap [q_\ell(r, \mathbf{v}), q_u(r, \mathbf{v})] \neq \emptyset\}.$$

Next, we describe the concept of variability. Before that we introduce some helpful related notions. Let η be a joint distribution over $\text{PO}_k \times \text{PO}_k$. We interpret it as a joint distribution on a, b:

$$\eta_{xy} \stackrel{def}{=} \Pr[a = x, b = y].$$

The marginal distribution denoted by η_a is defined as: $\forall x \in \text{PO}_k : \eta_a(x) \stackrel{def}{=} \sum_{y \in \text{PO}_k} \eta_{xy}$. η_b is define similarly. Let Δ_k be the simplex of all probability distributions over PO_k:

$$\Delta_k \stackrel{def}{=} \left\{ (g_0, g_1, \ldots, g_{k-1}) \mid g_i \geq 0 \text{ for all } i, \text{ and } \sum_{i=0}^{k-1} g_i = 1 \right\}.$$

Definition 8 (Lower Quantile) *Given a joint distribution η, the lower quantile of a particular value $r \in \{-(k-1), \cdots, k-1\}$ in the range of ITE, is defined as the total probability mass of all possible $(x, y) \in \text{PO}_k \times \text{PO}_k$ that have an ITE value strictly less than r. Formally, we have:*

$$q_\ell(r, \eta) \stackrel{def}{=} \sum_{x \in \text{PO}_k, y \in \text{PO}_k} \mathbf{1}\{x - y < r\} \eta_{xy}.$$

Definition 9 (Upper Quantile) *Given a joint distribution η, the upper quantile of a particular value $r \in \{-(k-1), \cdots, k-1\}$ in the range of ITE, is defined as the total probability mass of all possible $(x, y) \in \text{PO}_k \times \text{PO}_k$ that have an ITE value of at most r. Formally, we have:*

$$q_u(r, \eta) \stackrel{def}{=} \sum_{x \in \text{PO}_k, y \in \text{PO}_k} \mathbf{1}\{x - y \leq r\} \eta_{xy}.$$

Next, we prove a proposition which is useful when we want to translate between the lower qunatile of a vector and a lower qunatile of a distribution. It roughly states that for a typical sample $(\mathbf{a}, \mathbf{b})$ from $\eta^{\otimes n}$ (where η is a joint distribution over $\text{PO}_k \times \text{PO}_k$) we have $q_\ell(r, \eta) = q_\ell(r, \mathbf{a} - \mathbf{b})$ and $q_u(r, \eta) = q_u(r, \mathbf{a} - \mathbf{b})$. Further, $(\mathbf{a}, \mathbf{b}) \in \text{PO}_k^n \times \text{PO}_k^n$ there is a joint distribution η such that $(\mathbf{a}, \mathbf{b})$ is typical for $\eta^{\otimes n}$.

Proposition 10 *Let η be a joint distribution over $\text{PO}_k \times \text{PO}_k$. A sample $(\mathbf{a}, \mathbf{b})$ from $\eta^{\otimes n}$ is said to be typical if for all $(x, y) \in \text{PO}_k \times \text{PO}_k$ we have that the empirical frequency of $\rho(x, y) = \frac{|i \in [n] : (a_i, b_i) = (x, y)|}{n}$ is η_{xy}. Suppose $\mathbf{a}, \mathbf{b}$ is a typical sample from $\eta^{\otimes n}$, then, for all $r \in \mathbb{R}$ we have $q_\ell(r, \eta) = q_\ell(r, \mathbf{a} - \mathbf{b})$ and $q_u(r, \eta) = q_u(r, \mathbf{a} - \mathbf{b})$. Further, for any $(\mathbf{a}, \mathbf{b}) \in \text{PO}_k^n \times \text{PO}_k^n$ there is a joint distribution η such that $(\mathbf{a}, \mathbf{b})$ is typical for $\eta^{\otimes n}$.*

Proof Suppose $\mathbf{a}, \mathbf{b}$ is a typical sample from $\eta^{\otimes n}$. Let $\mathbf{v} = \mathbf{a} - \mathbf{b}$. Then,

$$q_\ell(r, \mathbf{v}) = \frac{|\{i \in [n] \mid \mathbf{v}_i < r\}|}{n} = \sum_{x \in \mathrm{PO}_k, y \in \mathrm{PO}_k} \mathbf{1}\{x - y < r\}\rho_{xy} = \sum_{x \in \mathrm{PO}_k, y \in \mathrm{PO}_k} \mathbf{1}\{x - y < r\}\eta_{xy} = q_\ell(r, \eta).$$

A similar argument holds for $q_u(r, \eta) = q_u(r, \mathbf{a} - \mathbf{b})$. Next, given any $(\mathbf{a}, \mathbf{b}) \in \mathrm{PO}_k^n \times \mathrm{PO}_k^n$ we associate a joint distribution η over $\mathrm{PO}_k \times \mathrm{PO}_k$ as follows: sample $i \in [n]$ uniformly at random and output (a_i, b_i). It is easy to see that $(\mathbf{a}, \mathbf{b})$ is typical for $\eta^{\otimes n}$. ■

Next, we define the concept of variability which captures how much the quantiles of a given estimate r can vary across all joint distributions that explain the given marginals. Given marginal distributions η_a and η_b, we define $\mathcal{J}(\eta_a, \eta_b)$ as the set of joint distributions η with marginals η_a and η_b for the potential outcomes.

Definition 11 (Variability) *Variability is a function $\nu : \mathbb{R} \times \Delta_k \times \Delta_k \to [0, 1] \times [0, 1]$, that takes as input an estimate r and marginal distributions for the potential outcomes η_a, η_b, and returns a tuple as follows:*

$$\nu(r, \eta_a, \eta_b) \stackrel{def}{=} (\nu_l(r, \eta_a, \eta_b), \nu_u(r, \eta_a, \eta_b)), \text{ where:}$$

$$\nu_l(r, \eta_a, \eta_b) = \max_{\eta \in \mathcal{J}(\eta_a, \eta_b)} q_l(r, \eta) \text{ and } \nu_u(r, \eta_a, \eta_b) = \min_{\eta \in \mathcal{J}(\eta_a, \eta_b)} q_u(r, \eta).$$

Next, we understand that given the variability of an estimate r, how far can it be from the median in a quantile sense.

Definition 12 (Minimum Median Width of Marginals) *Let η_a and η_b be distributions over PO_k. Further, for any $r \in \mathbb{R}$ the* width-of-r *for the pair (η_a, η_b) is defined as*

$$\varepsilon(r, \eta_a, \eta_b) := \max\{(\nu_l(r, \eta_a, \eta_b) - 1/2)^+, (1/2 - \nu_u(r, \eta_a, \eta_b))^+\}.$$

Then, the minimum median width *of the pair (η_a, η_b), denoted by $\epsilon^*(\eta_a, \eta_b)$, is defined as*

$$\min_{r \in \{-(k-1), \cdots, (k-1)\}} \varepsilon(r, \eta_a, \eta_b).$$

Remark 13 *Notice that for a typical sample $(\mathbf{a}, \mathbf{b})$ from $\eta^{\otimes n}$ such that the marginals of η are η_a and η_b we have that $r \in (\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}$ where $\epsilon = \varepsilon(r, \eta_a, \eta_b)$.*

2.1. Formalizing the Algorithm

Next, we formalize the meaning of an algorithm which respects the inference constraints of observing either a_i or b_i but not both. At the end of observing the vectors a and b in a random manner (even in an adaptive manner) it outputs a member m and a parameter $\epsilon \in [0, 0.5]$ and the output is correct if $m \in \left(\mathbf{a} - \mathbf{b}\right)_{0.5-\epsilon}^{0.5+\epsilon}$.

We first formalize the case of a deterministic algorithm, viewed as a decision tree.

Definition 14 (Feasible Decision Tree (FDT)) *A* feasible decision tree *(FDT) for block-length n is a rooted and labelled k-ary tree of depth n. Each non-leaf node is labelled with some element from the set $\{a, b\} \times [n]$ and each leaf node is labelled with an element from $\mathbb{R} \times [0, 0.5]$. The edges of the tree are labelled with elements from PO_k such that from any non-leaf node the set of labels on the edges to its children is PO_k. Finally, the tree should also satisfy the following* feasibility constraint. *For any path P from the root to some leaf and for all $i \in [n]$:*

$$|\{\mathrm{lab}(u) \mid u \in P \text{ and is a non-leaf node}\} \cap (\{a, b\} \times \{i\})| = 1.$$

Here, $\mathrm{lab}(u)$ refers to the label of a vertex u.

Definition 15 (Output of an FDT) *Let A_n be an FDT for block-length n. On input $(\mathbf{a}, \mathbf{b}) \in \mathrm{PO}_k^n \times \mathrm{PO}_k^n$ the output of A_n, denoted as $A_n(\mathbf{a}, \mathbf{b})$, is the label of the unique leaf obtained by following the path from the root to the leaf, where at each non-leaf node u, labelled (c, i), we follow the edge from u labelled with the same value as c_i: for instance, if $c = a$ and $a_i = 0$, then we follow the edge from u which is labeled 0. So, the output is a tuple $(m, \epsilon) \in \mathbb{R} \times [0, 0.5]$.*

Next, we define the case of a randomized algorithm.

Definition 16 (Randomized FDT (RFDT)) *A randomized feasible decision tree (RFDT) for block-length n is a distribution over the set of all FDT's for block-length n.*

Remark 17 *Definition 16 is equivalent to the description of RFDTs given in section 1.1.1. To see this, we can imagine pre-sampling all the internal randomness as required in the description of an RFDT in Section 1.1.1. In particular, for all non-leaf nodes u of the RFDT we fix a sample according to γ_u before proceeding along the RFDT. Given the samples from γ_u for each u, the RFDT is now actually an FDT. Hence, overall the RFDT is actually a distribution over FDTs.*

Definition 18 (Output of an RFDT) *Let R_n be an RFDT for block-length n and let μ_n be its distribution over the set of FDT's. On input $(\mathbf{a}, \mathbf{b}) \in \mathrm{PO}_k^n \times \mathrm{PO}_k^n$ the output of R_n, denoted as $R_n(\mathbf{a}, \mathbf{b})$, is the random variable obtained in the following manner. Sample an FDT A_n according to μ_n and output $A_n(\mathbf{a}, \mathbf{b})$.*

Next, we define the notions of width and error of an RFDT.

Definition 19 (Error and Width of an RFDT) *Fix $\delta \in [0, 1]$. An RFDT, R_n, for block-length n, is said to be a median approximator with error δ or a δ-error median approximator, if for all $(\mathbf{a}, \mathbf{b}) \in \mathrm{PO}_k^n \times \mathrm{PO}_k^n$:*

$$\Pr_{(m,\epsilon) \leftarrow R_n(\mathbf{a},\mathbf{b})} \left[m \in (a - b)_{0.5-\epsilon}^{0.5+\epsilon} \right] \geq 1 - \delta.$$

Further, the width of R_n on input $(\mathbf{a}, \mathbf{b})$, denoted as $\epsilon^{R_n}(\mathbf{a}, \mathbf{b})$, is defined as $\mathop{\mathbb{E}}_{(m,\epsilon) \leftarrow R_n(\mathbf{a},\mathbf{b})}[\epsilon]$.

3. Variability & Minimum Median Width

We first prove Lemma 20 which roughly states that any RFDT is incapable of distinguishing between two joint distributions over $\mathrm{PO}_k \times \mathrm{PO}_k$, say η and τ, which have the same marginals, i.e., $\eta_a = \tau_a$ and $\eta_b = \tau_b$. Specifically, any RFDT R_n is unable to distinguish whether the input is drawn from the distribution $\eta^{\otimes n}$ or $\tau^{\otimes n}$.

The proof proceeds by noting than any RFDT R_n is a distribution over FDTs. Further, for any FDT A_n and for any path P in A_n the set of inputs $(\mathbf{a}, \mathbf{b})$ where A_n follows the path P, denoted by $F(A_n, P)$, is described by constraints on exactly one of a_i or b_i for $i \in [n]$. Hence, $\Pr_{(\mathbf{a},\mathbf{b})\leftarrow\eta^{\otimes n}}[(\mathbf{a}, \mathbf{b}) \in F(A_n, P)]$ is only a function of the marginals η_a, η_b. Thus, the probability of $F(A_n, P)$ remains the same under τ. Proof appears in Section A.

Lemma 20 (Limits of RFDT) *Let η, τ be joint distributions over $\mathrm{PO}_k \times \mathrm{PO}_k$ with same marginals for a, b, i.e., $\eta_a = \tau_a$ and $\eta_b = \tau_b$, and let R_n be an RFDT for block-length n. Further, let $R_n(\eta), R_n(\tau)$ be the distributions of the output of R_n when the input $(\mathbf{a}, \mathbf{b})$ is sampled according to $\eta^{\otimes n}$ and $\tau^{\otimes n}$ respectively. Then, $R_n(\eta) = R_n(\tau)$. More precisely, for every $(m', \epsilon') \in \mathbb{R} \times [0, 0.5]$ we have*

$$\Pr_{\substack{(\mathbf{a},\mathbf{b})\leftarrow\eta^{\otimes n}\\(m,\epsilon)\leftarrow R_n(a,b)}}[(m, \epsilon) = (m', \epsilon')] = \Pr_{\substack{(\mathbf{a},\mathbf{b})\leftarrow\tau^{\otimes n}\\(m,\epsilon)\leftarrow R_n(a,b)}}[(m, \epsilon) = (m', \epsilon')].$$

Using Lemma 20 we prove the theorem below which serves to capture the hardness of approximation the MTE in terms of the variability of the marginal distributions of the treatment and control vector. As a consequence, if we can devise an algorithm for computing $\epsilon^*(\eta_a, \eta_b)$ exactly (section Section 4), then, then our algorithm is essentially optimal.

We describe the proof idea after the formal statement (refer to Appendix B for full proof).

Theorem 21 (Lower Bound on Width of any RFDT) *Let η be any joint distribution on $\mathrm{PO}_k \times \mathrm{PO}_k$ with marginals η_a and η_b for the treatment and control respectively. Further, let R_n be a δ-error median approximator for block-length n. Then, for any $\beta > 0$ we have*

$$\mathop{\mathbb{E}}_{(\mathbf{a},\mathbf{b})\leftarrow\eta^{\otimes n}}[\epsilon^{R_n}(\mathbf{a}, \mathbf{b})] \geq (\epsilon^*(\eta_a, \eta_b) - \beta) \cdot (1 - 2k\delta) + 2k \exp(-2\beta^2 \cdot n).$$

The proof idea is as follows. Fix a a δ-error median approximator R_n. Given a joint distribution η which has marginals η_a and η_b, for an estimate $r \in \{-(k-1), \ldots, k-1\}$, let γ_r denote the probability that R_n outputs $m = r$ and $\epsilon < \epsilon(r, \eta_a, \eta_b)$. Now, consider the joint distribution $\eta^{(r)}$ which witnesses the width-of-r value, i.e.,such that at least one of $\epsilon(r) = \left(q_\ell(r, \eta^{(r)}) - 1/2\right)^+$ or $\epsilon(r) = (1/2 - q_u(r, \eta^{(r)}))^+$ is true. For $\eta^{(r)}$ the probability that R_n outputs $m = r$ and $\epsilon < \epsilon(r, \eta_a, \eta_b)$ is still γ_r. However if $\epsilon < \epsilon(r, \eta_a, \eta_b)$, then, for a typical sample from $\left(\eta^{(r)}\right)^{\otimes n}$ it will be the case that $r \notin (\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}$, and this counts as an error. Hence, γ_r is not too much larger than δ. Overall, taking a union over all values for r, we get that except for around a $2k\delta$ probability the output of R_n has to be at least $\epsilon^*(\eta_a, \eta_b)$.

3.1. Bounds on Variability

In this section, we show that the value of $\epsilon^*(\eta_a, \eta_b)$ is upper bounded by $\frac{1}{2} \cdot \frac{(2k-3)}{(2k-1)}$. In doing so, we have essentially shown that any algorithm that computes the variability exactly obtains an error that is upper bounded by $\frac{1}{2} \cdot \frac{(2k-3)}{(2k-1)}$. The proof of the lemma below is provided in Section B.2.

Algorithm 1 MEDIAN-ESTIMATOR

Input: Treatment and control response vectors $\widehat{\mathbf{a}}, \widehat{\mathbf{b}}$ and slack parameter $\beta > 0$
Output: Median estimate $\widehat{m}$ and the width ϵ
Compute the empirical distributions ρ_a, ρ_b as follows:

$$\forall j \in \text{PO}_k \quad \rho_a(j) = \frac{|\{i|\widehat{\mathbf{a}}_i = j \text{ for } i \in \{1, 2, \cdots, n\}\}|}{n/2}$$
$$\forall j \in \text{PO}_k \quad \rho_b(j) = \frac{|\{i|\widehat{\mathbf{b}}_i = j \text{ for } i \in \{1, 2, \cdots, n\}\}|}{n/2}$$

for r in $\{-(k-1), \ldots, (k-1)\}$ **do**
  Compute $\nu_\ell(r, \rho_a, \rho_b)$ and $\nu_u(r, \rho_a, \rho_b)$ using Algorithms 2 and 3
  Compute $\varepsilon(r, \rho_a, \rho_b) := \max\left\{\left(\nu_l(r, \rho_a, \rho_b) - \frac{1}{2}\right)^+, \left(\frac{1}{2} - \nu_u(r, \rho_a, \rho_b)\right)^+\right\}$
end for
$\widehat{m} \leftarrow \underset{r \in \{-(k-1), \ldots, (k-1)\}}{\text{argmin}} \varepsilon(r, \rho_a, \rho_b)$; $\epsilon \leftarrow \varepsilon(\widehat{m}, \rho_a, \rho_b) + 2k\beta$.
return $\widehat{m}, \epsilon$.

Lemma 22 (Upper Bound on Minimum Median Width) *Given two distributions η_a and η_b over* PO_k*, we have* $\epsilon^*(\eta_a, \eta_b) \leq \frac{1}{2} \cdot \frac{(2k-3)}{(2k-1)}$.

We show a tightness result by arguing that there exists η_a, η_b that satisfy the stated upper bound. The proof appears in Section B.2.

Theorem 23 (Extremal distributions for Minimum Median Width) *For any $k \in \mathcal{N}$, $r \in \text{PO}_k$, let η_a, η_b be distributions over* PO_k *defined as:*

$$\eta_a(r) = \frac{1}{2k-1} \cdot (1 + \mathbf{1}\{r > 0\}) \text{ and } \eta_b(r) = \frac{1}{2k-1} \cdot (1 + \mathbf{1}\{r < k-1\}).$$

Then, $\epsilon^*(\eta_a, \eta_b) \geq \frac{1}{2} \cdot \frac{2k-3}{2k-1}$. Combined with Lemma 22 we have $\epsilon^*(\eta_a, \eta_b) = \frac{1}{2} \cdot \frac{2k-3}{2k-1}$.

4. Algorithm

In this section, we describe the Algorithm MEDIAN-ESTIMATOR that takes as input the observed response vectors $\widehat{\mathbf{a}}$, $\widehat{\mathbf{b}}$, and returns an estimate $\widehat{m}$ for the median of the vector $\mathbf{a} - \mathbf{b}$, along with an error estimate ϵ; with a guarantee that $\widehat{m}$ is within an ϵ band around the median, i.e., $\widehat{m} \in (\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}$. The missing proofs are collected in Appendix C and D.

First, in section 4.1, we describe an algorithm MEDIAN-ESTIMATOR that uses the algorithm COMPUTE-VARIABILITY and returns a median estimate $\widehat{m}$ and its ϵ band. Next, in section 4.2 we describe an algorithm COMPUTE-VARIABILITY that computes variability $\nu(r, \eta_a, \eta_b)$, given a feasible median value $r \in \{-(k-1), \cdots, (k-1)\}$, and marginals η_a, η_b.

4.1. Approximating the Median

In Algorithm 1, we first compute the empirical distributions ρ_a, ρ_b that are close approximations to η_a, η_b (up to Chernoff errors). Then, using Algorithms COMPUTE-VARIABILITY-LOWER-QUANTILE and COMPUTE-VARIABILITY-UPPER-QUANTILE (in Appendix D), we compute the variability for every feasible median value in the set $\{-(k-1), \cdots, (k-1)\}$. The minimizer in the set is output the median estimate $\widehat{m}$ and the corresponding minimum median width (see Definition 12) $\epsilon(\widehat{m}, \rho_a, \rho_b)$ as the approximation error.

Theorem 24 *Let $\mathbf{a}, \mathbf{b} \in \mathrm{PO}_k^n$ be the treatment and control vectors. Define the true empirical frequency vectors as $\eta_a(j) = \frac{|\{i \in [n] | \mathbf{a}_i = j\}|}{n}$ and $\eta_b(j) = \frac{|\{i \in [n] | \mathbf{b}_i = j\}|}{n}$. Suppose that the* Bernoulli Design *is used to assign the individuals to the treatment and control groups, i.e., each individual is assigned to one of the groups independently with probability $1/2$. Let the observed response vectors be $\widehat{\mathbf{a}}, \widehat{\mathbf{b}}$. Then, Algorithm* MEDIAN-ESTIMATOR *is a δ-error median estimator $\widehat{m}$ (Definition 19), with width $\epsilon \leq \epsilon^*(\eta_a, \eta_b) + 2k\beta$, where $\delta = 2k \cdot \exp\left(-2\beta^2 n\right)$ and runs in $O(n + k^2)$ time.*

Note. As k increases, the approximation error ϵ approaches $1/2$, which is a vacuous bound. However, this is inescapable owing to an almost matching lower-bound. From Theorem 24 for all $n = \omega(k^2)$, the value ϵ returned by our algorithms is non-trivial, ie, less than $1/2$. The bottleneck at $n \approx k^2$ happens due to the errors from estimating the marginals. In other words, when $n \ll k^2$, our estimated marginals have high approximation error (from Chernoff bounds). We note that our algorithms running time of $O(n + k^2)$ is optimal with respect to n. This is because $\Omega(n)$ running time is necessary to even read all the n potential outcome values, i.e., vectors $\mathbf{a}, \mathbf{b}$. When $k = o(\sqrt{n})$, the running time is $O(n)$, which is optimal with respect to n.

4.2. Computing Variability

In section 3, we have shown that computing the minimum median width of marginals (by calculating the variability for every feasible median value) captures the inherent complexity of approximating the median. In this section, we show that there is a natural greedy algorithm that computes the variability exactly. To compute variability, we need to compute $\nu_l(r, \eta_a, \eta_b)$ and $\nu_u(r, \eta_a, \eta_b)$, where r is a feasible median value and η_a, η_b are the marginals. As $\nu_l(r, \eta_a, \eta_b) = \max_\eta \sum_{x,y} \mathbf{1}\{x - y < r\}\ \eta_{xy}$, we need to identify an appropriate joint distribution η, that maximizes the total probability for the lower quantile. It turns out that we can formulate this as a linear program shown below (see Figure 1), with marginal and non-negativity constraints.

We can view solving the linear program (in Figure 1) as essentially filling in the entries of a 2-dimensional matrix of size $\mathrm{PO}_k \times \mathrm{PO}_k$ such that the difference in indices is strictly smaller than r, i.e., $x - y < r\ \forall x, y$. To do so, we employ a greedy strategy in Algorithm COMPUTE-VARIABILITY-LOWER-QUANTILE (see Alg 2). Without loss of generality, we fill the matrix, greedily from the last column and last row and move up the column (see Lines 4-5 in Alg 2). Our greedy strategy involves increasing the value of η_{xy} until one of the marginal constraints is tight and freezing the values of the row/column corresponding to the constraint (see Lines 7-10 in Alg 2). A similar approach results in an algorithm 3 for computing the upper quantile component of variability. Formally:

Lemma 25 *Given a feasible value $r \in \{-(k-1), \cdots, (k-1)\}$ and marginals η_a, η_b, there are algorithms for computing the values $\nu_l(r, \eta_a, \eta_b)$ and $\nu_u(r, \eta_a, \eta_b)$ exactly, and run in $O(k)$ time.*

$$\begin{aligned}
\underline{\text{Objective}}: \quad & \max \sum_{x,y} \mathbf{1}\{x - y < r\}\eta_{xy} \\
\text{(Marginal Constraints)} \quad & \sum_{y \in \text{PO}_k} \eta_{xy} = \eta_a(x) \quad \forall x \in \text{PO}_k \\
\text{(Marginal Constraints)} \quad & \sum_{x \in \text{PO}_k} \eta_{xy} = \eta_b(y) \quad \forall y \in \text{PO}_k \\
\text{(Non-negativity)} \quad & \eta_{xy} \geq 0 \quad \forall x, y \in \text{PO}_k \times \text{PO}_k
\end{aligned}$$

Figure 1: Linear Program for computing the lower quantile component of variability

Adapting our algorithm to designs other than the Bernoulli design. For our approach, the estimated empirical marginals ρ_a, ρ_b should have low approximation errors with respect to the exact marginals. For designs that satisfy this requirement, e.g., balanced design, or Bernoulli design with a different probability of assignment, we can directly extend our approach.

5. Conclusion

In this work, we studied the task of median treatment effect estimation, to capture the heterogeneity in the data. We argued that the task is inestimable, provided new notions of approximations, and showed optimal algorithms with minimum approximation error. We want to highlight that our results make no distributional assumptions, and can be extended for any quantile $q < 0.5$ (in addition to median), by modifying the definition of variability (definition 11) appropriately. For future work, a potential direction is to study the settings where the potential outcomes are continuous. We note that our algorithms are optimal with respect to the approximation parameters, and not necessarily with respect to computational complexity (running time). When $k = o(\sqrt{n})$, the running time is $O(n)$, is optimal with respect to n. For other regimes, the term k^2 need not necessarily be optimal, and is an interesting future direction. Another interesting future direction is to incorporate a competing objective, such as (Harshaw et al., 2023) with respect to balancing the covariates.

Acknowledgments

The authors would like to thank the following individuals for many helpful discussions: Anup Rao, Kirankumar Shiragur, Christopher Harshaw, Shiva Kasiviswanathan, and Avi Feller. The authors also would like to thank the organizers for the invitation to the Causality program at the Simons Institute for the Theory of Computing, which helped facilitate early discussions of this work.

References

Raghavendra Addanki, David Arbour, Tung Mai, Cameron Musco, and Anup Rao. Sample constrained treatment effect estimation. *Advances in Neural Information Processing Systems*, 35: 5417–5430, 2022.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Iavor Bojinov, Guillaume Saint-Jacques, and Martin Tingley. Avoid the pitfalls of a/b testing make sure your experiments recognize customers' varying needs. *Harvard Business Review*, 98(2): 48–53, 2020.

Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Peng Ding, Avi Feller, and Luke Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

Christopher Harshaw, Fredrik Sävje, Daniel A Spielman, and Peng Zhang. Balancing covariates in randomized experiments with the gram–schmidt walk design. *Journal of the American Statistical Association*, pages 1–27, 2023.

James J Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

Steven R Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to a/b testing and best-arm identification. *Bernoulli*, 28(3):1704–1728, 2022.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, 7(1):443–470, 2013.

Nathan Kallus. What's the harm? sharp bounds on the fraction negatively affected by treatment. *Advances in Neural Information Processing Systems*, 35:15996–16009, 2022.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

Liu Leqi and Edward H Kennedy. Median optimal treatment regimes. *arXiv preprint arXiv:2103.01802*, 2021.

Diego Martinez-Taboada and Edward H Kennedy. Counterfactual density estimation using kernel stein discrepancies. *arXiv preprint arXiv:2309.16129*, 2023.

Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni. Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

Appendix A. Limitations of RFDTs

A.0.1. PROOF OF LEMMA 20

Proof Fix some (m', ϵ'). Let $\mathrm{supp}(R_n)$ denote the support of the distribution of R_n, i.e., the support of the distribution over FDTs of block-length n which represents R_n. Fix any FDT $A_n \in \mathrm{supp}(R_n)$ and any path from the root to leaf in A_n, say P, such that the label of the leaf node is (m', ϵ'). Further, let $F(A_n, P)$ denote those inputs, $(\mathbf{a}, \mathbf{b}) \in \mathrm{PO}_k^n \times \mathrm{PO}_k^n$ such that A_n follows the path P on the input $(\mathbf{a}, \mathbf{b})$. We will show that

$$\Pr_{\mathbf{a},\mathbf{b} \leftarrow \eta^{\otimes n}}[F(A_n, P)] = \Pr_{\mathbf{a},\mathbf{b} \leftarrow \tau^{\otimes n}}[F(A_n, P)].$$

To see this, note that there is a function $f : [n] \to \mathrm{PO}_k$ and a subset $S \subseteq [n]$ such that $F(A_n, P) = \{(\mathbf{a}, \mathbf{b}) \mid a_i = f(i) \text{ for } i \in S, b_i = f(i) \text{ for } i \notin S\}$. This follows from the feasibility constraint on A_n as mentioned in Definition 14. Hence, we have

$$\begin{aligned}\Pr_{\mathbf{a},\mathbf{b} \leftarrow \eta^{\otimes n}}[F(A_n, P)] &= \prod_{i \in S} \eta_a(f(i)) \times \prod_{i \notin S} \eta_b(f(i)) \\ &= \prod_{i \in S} \tau_a(f(i)) \times \prod_{i \notin S} \tau_b(f(i)) = \Pr_{\mathbf{a},\mathbf{b} \leftarrow \tau^{\otimes n}}[F(A_n, P)].\end{aligned}$$

Taking a union over all paths P in A_n whose leaves are labelled with (m', ϵ') it follows that

$$\Pr_{\mathbf{a},\mathbf{b} \leftarrow \eta^{\otimes n}}[A_n((\mathbf{a}, \mathbf{b})) = (m', \epsilon')] = \Pr_{\mathbf{a},\mathbf{b} \leftarrow \tau^{\otimes n}}[A_n((\mathbf{a}, \mathbf{b})) = (m', \epsilon')].$$

Since, the above statement is true for every fixed A_n in the support of R_n the lemma follows. ■

Appendix B. Variabilty

B.1. Lower Bounds on Width

B.1.1. PROOF OF THEOREM 21

Proof For $r \in \{-(k-1), \ldots, k-1\}$ recall that (see width-of-r in Definition 12)

$$\varepsilon(r, \eta_a, \eta_b) = \max\{(\nu_\ell(r, \eta_a, \eta_b) - 1/2)^+, (1/2 - \nu_u(r, \eta_a, \eta_b))^+\},$$

where ν_ℓ and ν_u are as defined in Definition 11. For the remainder of the proof we will use $\epsilon(r)$ and ϵ^* to denote $\epsilon(r, \eta_a, \eta_b)$ and $\epsilon^*(\eta_a, \eta_b)$ respectively. Note that $\epsilon^* = \min_r \epsilon(r)$.

Further, let $\eta^{(r)}$ denote a joint distribution over $\mathrm{PO}_k \times \mathrm{PO}_k$ with marginals η_a, η_b such that at least one of $\epsilon(r) = \left(q_\ell(r, \eta^{(r)}) - 1/2\right)^+$ or $\epsilon(r) = (1/2 - q_u(r, \eta^{(r)}))^+$ is true. Now, let

$$\gamma_r = \Pr_{\substack{(\mathbf{a},\mathbf{b}) \leftarrow \eta^{\otimes n} \\ (m,\epsilon) \leftarrow R_n(\mathbf{a},\mathbf{b})}}[m = r, \epsilon < \epsilon(r) - \beta].$$

Notice that by Lemma 20 we have

$$\gamma_r = \Pr_{\substack{(\mathbf{a},\mathbf{b}) \leftarrow (\eta^{(r)})^{\otimes n} \\ (m,\epsilon) \leftarrow R_n(\mathbf{a},\mathbf{b})}}[m = r, \epsilon < \epsilon(r) - \beta],$$

i.e., the value of γ_r remains unchanged when inputs are sampled from $\left(\eta^{(r)}\right)^{\otimes n}$ instead of $\eta^{\otimes n}$.

Now, for an input $(\mathbf{a}, \mathbf{b})$ sampled according to $\left(\eta^{(r)}\right)^{\otimes n}$ we have with at least $1-\exp(-2\beta^2 \cdot n)$ probability that $r \notin (\mathbf{a}-\mathbf{b})_{0.5-\epsilon(r)+\beta}^{0.5+\epsilon(r)-\beta}$. This follows from the Chernoff bound. Suppose $\epsilon(r) = \left(q_\ell(r, \eta^{(r)}) - 1/2\right)$, and note that with probability at least $1-\exp(-2\beta^2 \cdot n)$ the empirical frequency of ITE's less than r will be at least $q_\ell(r, \eta^{(r)}) - \beta$, and hence $q_\ell(r, \mathbf{a}-\mathbf{b}) \geq q_\ell(r, \eta^{(r)}) - \beta$ with probability at least $1 - \exp(-2\beta^2 \cdot n)$.

Notice that this yields $\gamma_r \leq \delta + \exp(-2\beta^2 \cdot n)$ as

$$\begin{aligned}
\gamma_r &= \Pr_{\substack{(\mathbf{a},\mathbf{b})\leftarrow\left(\eta^{(r)}\right)^{\otimes n} \\ (m,\epsilon)\leftarrow R_n(\mathbf{a},\mathbf{b})}}[m = r, \epsilon < \epsilon(r) - \beta] \\
&\leq \Pr_{\substack{(\mathbf{a},\mathbf{b})\leftarrow\left(\eta^{(r)}\right)^{\otimes n} \\ (m,\epsilon)\leftarrow R_n(\mathbf{a},\mathbf{b})}}[m = r, \epsilon < \epsilon(r) - \beta, r \notin (\mathbf{a}-\mathbf{b})_{0.5-\epsilon(r)+\beta}^{0.5+\epsilon(r)-\beta}] + \exp(-2\beta^2 \cdot n) \\
&\leq \Pr_{\substack{(\mathbf{a},\mathbf{b})\leftarrow\left(\eta^{(r)}\right)^{\otimes n} \\ (m,\epsilon)\leftarrow R_n(\mathbf{a},\mathbf{b})}}[m \notin (\mathbf{a}-\mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}] + \exp(-2\beta^2 \cdot n) \leq \delta + \exp(-2\beta^2 \cdot n).
\end{aligned}$$

Thus,

$$\mathop{\mathbb{E}}_{(\mathbf{a},\mathbf{b})\leftarrow\eta^{\otimes n}}[\epsilon^{R_n}(\mathbf{a},\mathbf{b})] \geq (\epsilon^* - \beta) \times \Pr_{\substack{(\mathbf{a},\mathbf{b})\leftarrow\eta^{\otimes n} \\ (m,\epsilon)\leftarrow R_n(\mathbf{a},\mathbf{b})}}[\epsilon \geq \epsilon^* - \beta],$$

which by a union bound over all $r \in \text{ITE}$ gives

$$\begin{aligned}
&\geq (\epsilon^* - \beta) \times \left(1 - \sum_{r\in\{-(k-1),\ldots,k-1\}} \gamma_r\right) \\
&\geq (\epsilon^*(\eta_a, \eta_b) - \beta) \cdot (1 - 2k\delta) + 2k\exp(-2\beta^2 \cdot n).
\end{aligned}$$

■

B.2. Bounds on Variability

B.2.1. PROOF OF LEMMA 22

For contradiction, let's assume that $\exists \eta_a, \eta_b$ such that $\epsilon^*(\eta_a, \eta_b) > \frac{1}{2} \cdot \frac{(2k-3)}{(2k-1)}$. For brevity let $\psi_k = \frac{1}{(2k-1)}$ and let $\epsilon(r, \eta_a, \eta_b)$, $\nu_\ell(r, \eta_a, \eta_b)$ and $\nu_u(r, \eta_a, \eta_b)$ be denoted as $\epsilon(r)$, $\nu_\ell(r)$ and $\nu_u(r)$ respectively. Therefore, we have:

$$\varepsilon(r) > 1/2 - \psi_k \quad \forall r \in \{-(k-1), \cdots, (k-1)\},$$

where from Definition 12

$$\varepsilon(r) = \max\{(\nu_l(r) - 1/2)^+, (1/2 - \nu_u(r))^+\}.$$

Notice that $\nu_\ell(-(k-1)) = 0$, thus, $\nu_u(-(k-1)) < \psi_k$. However, $\nu_u(k-1) = 1$, thus $\nu_\ell((k-1)) > 1-\psi_k$. Now, let $r \in \text{ITE}$ be the smallest value such that $\nu_\ell(r) > 1-\psi_k$, therefore, $\nu_\ell(r-1) \leq 1-\psi_k$ and hence $\nu_u(r-1) < \psi_k$. We show that the existence of such an r is impossible by analyzing $\mathbb{E}_{X\leftarrow\eta_a}[X] - \mathbb{E}_{Y\leftarrow\eta_b}[Y]$.

Notice that for any joint distribution η such that its marginals are η_a and η_b and $q_u(r-1,\eta) < \psi_k$ we have

$$\mathop{\mathbb{E}}_{X\leftarrow\eta_a}[X] - \mathop{\mathbb{E}}_{Y\leftarrow\eta_b}[Y] = \mathop{\mathbb{E}}_{X,Y\leftarrow\eta}[X-Y] > -(k-1)\cdot\psi_k + (r)\cdot(1-\psi_k).$$

Similarly for any joint distribution η' such that its marginals are η_a and η_b and $q_\ell(r,\eta) > 1-\psi_k$ we have

$$\mathop{\mathbb{E}}_{X\leftarrow\eta_a}[X] - \mathop{\mathbb{E}}_{Y\leftarrow\eta_b}[Y] = \mathop{\mathbb{E}}_{X,Y\leftarrow\eta'}[X-Y] < (r-1)\cdot(1-\psi_k) + (k-1)\cdot\psi_k.$$

Hence, we obtain

$$\begin{aligned} -(k-1)\cdot\psi_k + (r)\cdot(1-\psi_k) &< (r-1)\cdot(1-\psi_k) + (k-1)\cdot\psi_k \\ &\implies 1-\psi_k < 2(k-1)\psi_k \\ &\implies \frac{2k-2}{2k-1} < \frac{2k-2}{2k-1}, \end{aligned}$$

which is a contradiction. Therefore, the posited r can't exist and this completes the proof.

B.2.2. PROOF OF THEOREM 23

Proof We first show that $\nu_u(0,\eta_a,\eta_b) \leq \frac{1}{2k-1}$. For this consider the joint distribution η described as follows: sample $a \sim \eta_a$ and let $b = (a-1) \mod k$. It is easy to see that b is distributed as η_b. Further, $a-b$ is $-(k-1)$ with probability $\frac{1}{2k-1}$ and is 1 otherwise. Thus, $q_u(0,\eta) = \frac{1}{2k-1}$, and hence, $\nu_u(0,\eta_a,\eta_b) \leq \frac{1}{2k-1}$.

Next, we show that $\nu_\ell(1,\eta_a,\eta_b) \geq \frac{2k-2}{2k-1}$. For this consider the joint distribution η described as follows: sample $a \sim \eta_a$ and if $a \neq k-1$ let $b = a$; if $a = k-1$ then let $b = 0$ or $k-1$ with probability 0.5. It is easy to see that b is distributed as η_b. Further, $a-b$ is $(k-1)$ with probability $\frac{1}{2k-1}$ and is 0 otherwise. Thus, $q_\ell(1,\eta) = \frac{2k-2}{2k-1}$, and hence, $\nu_\ell(1,\eta_a,\eta_b) \geq \frac{2k-2}{2k-1}$.

Notice that both $\nu_u(r,\eta_a,\eta_b)$ and $\nu_\ell(r,\eta_a,\eta_b)$ are increasing functions of r. Hence, for all $r \leq 0$ we have $\nu_u(r,\eta_a,\eta_b) \leq \frac{1}{2k-1}$ and hence for all $r \leq 0$ we have $\epsilon(r,\eta_a,\eta_b) \geq 0.5\cdot\frac{2k-3}{2k-1}$. Similarly, for all $r \geq 1$ we have $\nu_\ell(r,\eta_a,\eta_b) \geq \frac{2k-2}{2k-1}$ and hence for all $r \geq 1$ we have $\epsilon(r,\eta_a,\eta_b) \geq 0.5\cdot\frac{2k-3}{2k-1}$. This proves the theorem. ■

Appendix C. Median Estimator

C.0.1. PROOF OF THEOREM 24

Proof Consider the observed empirical distributions ρ_a, ρ_b (Line 3 in Alg 1) obtained from the response vectors $\widehat{\mathbf{a}}, \widehat{\mathbf{b}}$. From Chernoff and union bound, we have the following:

$$\max_{j\in\text{PO}_k} \max\{|\rho_a(j) - \eta_a(j)|, |\rho_b(j) - \eta_b(j)|\} \leq \beta,$$

with probability at least $1 - 2k \cdot \exp\left(-2\beta^2 n\right)$.

Therefore, we focus on the case when the distributions ρ_a (or ρ_b) and η_a (or ρ_b) are separated by a point-wise distance of at most β; hence the total variation (TV distance between ρ_a (or ρ_b) and η_a (or ρ_b) is at most $(1/2)k\beta$. Our goal is to calculate the deviation in variability (and width) due to the this distance.

Consider the comparison of $\nu_\ell(r, \eta_a, \eta_b)$ and $\nu_\ell(r, \rho_a, \rho_b)$ for some r. Suppose η is the optimizer of the LP in Figure 1 for $\nu_\ell(r, \eta_a, \eta_b)$. Next, we show that there is a joint distribution ρ with marginals ρ_a and ρ_b which is within a TV distance $k\beta$ of η. Let $\mathcal{J}_a$ be an optimal coupling between η_a and ρ_a, i.e., a joint distribution with marginals η_a and ρ_a such that $\Pr_{X,X' \leftarrow \mathcal{J}_a}[X \neq X'] \leq (1/2)k\beta$. Similarly, define $\mathcal{J}_b$. Then, ρ is described as follows: sample $(X, Y) \leftarrow \eta$ and generate $X' \leftarrow \mathcal{J}_a|X$ and $Y' \leftarrow \mathcal{J}_b|Y$ independently, and output (X', Y'). It is easy to see that the distribution of (X', Y') is ρ and that $\Pr[(X, Y) \neq (X', Y')] \leq k\beta$. Therefore, $|\rho - \eta|_{TV} \leq k\beta$. However, this gives that $\Pr_{(X,Y)\leftarrow\eta}[X - Y < r]$ is $\Pr_{(X,Y)\leftarrow\eta}[X - Y < r] \pm k\beta$. Together with a symmetric argument starting with the optimizer of $\nu_\ell(r, \rho_a, \rho_b)$, we obtain $\nu_\ell(r, \eta_a, \eta_b)$ and $\nu_\ell(r, \rho_a, \rho_b) \pm \beta$. Hence, we also obtain $\epsilon^*(\eta_a, \eta_b)$ is $\epsilon^*(\rho_a, \rho_b) \pm k\beta$.

However, the above implies that the ϵ output by Algorithm 1 is at least $\epsilon^*(\eta_a, \eta_b)$ and at most $\epsilon^*(\eta_a, \eta_b) + 2k\beta$. Also, by Proposition 10 let η be the joint distribution such that $(\mathbf{a}, \mathbf{b})$ is typical for η. Then, we know that for each $r \in \text{ITE}$, $q_\ell(r, \mathbf{a} - \mathbf{b}) = q_\ell(r, \eta) \leq \nu_\ell(r, \eta_a, \eta_b) \leq \nu_\ell(r, \rho_a, \rho_b) + k\beta$ and $q_u(r, \mathbf{a} - \mathbf{b}) = q_u(r, \eta) \geq \nu_u(r, \eta_a, \eta_b) \geq \nu_u(r, \rho_a, \rho_b) - k\beta$. In particular, this is true for $r = m$, the output of Algorithm 1. Therefore, $m \in (\mathbf{a} - \mathbf{b})_{0.5-\epsilon}^{0.5+\epsilon}$. This completes the proof of correctness of Algorithm 1.

For the running time analysis observe that we calculate variability values for $2k - 1$ possible median values (Line 4 in Algorithm MEDIAN-ESTIMATOR) after computing the empirical distributions ρ_a, ρ_b. Therefore, our claim regarding the running time follows from Lemmas 26 and 27. ■

Appendix D. Computing Variability

Algorithm 2 COMPUTE-VARIABILITY-LOWER-QUANTILE

1: **Input:** Marginal distributions η_a, η_b and feasible median value $r \in \{-(k-1), \cdots, (k-1)\}$.
2: **Output:** $\nu_l(r, \eta_a, \eta_b)$.
3: $\forall x \in \text{PO}_k, y \in \text{PO}_k$, set $\eta_{xy} \leftarrow 0$
4: **for** $y \in \{k-1, k-2, \cdots, 0\}$ **do**
5: **for** $x \in \{y + r - 1, y + r - 2, \cdots, 0\} \cap \text{PO}_k$ **do**
6: Identify all the marginal constraints in LP 1 containing η_{xy} and then slowly increase the value of η_{xy} until one of them becomes tight.
7: We mark the corresponding column or row with the tight constraint as *frozen* and do not update any values in them.
8: **end for**
9: **end for**
10: **Return** $\nu_l(r, \eta_a, \eta_b) \leftarrow \sum_{x,y \in \text{PO}_k \times \text{PO}_k} \mathbf{1}\{x - y < r\}\eta_{xy}$.

Lemma 26 *Given a feasible median value $r \in \{-(k-1), \cdots, (k-1)\}$ and marginals η_a, η_b, Algorithm 2 computes the value $\nu_l(r, \eta_a, \eta_b)$ exactly and runs in time $O(k)$.*

Proof Let η^* denote the output of the greedy Algorithm and $\bar{\eta}$ be an optimizer of Figure 1. Let us assume:

$$\sum_{x,y} \mathbf{1}\{x - y < r\}\eta^*_{xy} < \sum_{x,y} \mathbf{1}\{x - y < r\}\bar{\eta}_{xy}$$

As the greedy algorithm fills the entries of η^*_{xy} by making sure that one of the constraints is tight: we will be able to identify a pair (wlog first pair) x, y such that $\eta^*_{xy} > \bar{\eta}_{xy}$. Otherwise, we would have $\bar{\eta} = \eta^*$. Now consider the marginal constraint

$$\sum_{x \in \text{PO}_k} \eta^*_{xy} = \sum_{x \in \text{PO}_k} \bar{\eta}_{xy} = \eta_b(y).$$

Therefore, we have a u such that: $\eta^*_{uy} < \bar{\eta}_{uy}$. Similarly the marginal constraint

$$\sum_{y \in \text{PO}_k} \eta^*_{xy} = \sum_{y \in \text{PO}_k} \bar{\eta}_{xy} = \eta_a(x).$$

Therefore, we have a v such that: $\eta^*_{xv} < \bar{\eta}_{xv}$. We update $\bar{\eta}$ by increasing the value of $\bar{\eta}_{xy}$ as follows:

1. $\bar{\eta}^{\text{upd}}_{uy} = \max\{0, \bar{\eta}_{uy} - (\eta^*_{xy} - \bar{\eta}_{xy})\}$
2. $\bar{\eta}^{\text{upd}}_{xv} = \max\{0, \bar{\eta}_{xv} - (\eta^*_{xy} - \bar{\eta}_{xy})\}$
3. $\bar{\eta}^{\text{upd}}_{uv} = \bar{\eta}_{uv} + (\eta^*_{xy} - \bar{\eta}_{xy})$
4. $\bar{\eta}^{\text{upd}}_{xy} = \bar{\eta}_{xy} + (\eta^*_{xy} - \bar{\eta}_{xy})$

Observe that such an update satisfies the marginal constraints and the objective function remains the same for both $\bar{\eta}$ and $\bar{\eta}^{\text{upd}}$. The updated $\bar{\eta}^{\text{upd}}$ such that $\bar{\eta}^{\text{upd}}_{xy} = \eta^*_{xy}$. By repeatedly following this updating process, we will have $\bar{\eta}^{\text{upd}} = \eta^*$, a contradiction to our claim. As we eliminate a row or column by making a marginal constraint tight in each iteration of our algorithm. The total number of iterations and the values of $(x, y) \in \text{PO}_k \times \text{PO}_k$ pairs that we fill are at most $2k$. Hence, the running time of our algorithm is $O(k)$. ■

As $\nu_u(r, \eta_a, \eta_b) = \min_\eta \sum_{x,y} \mathbf{1}\{x - y \leq r\}\ \eta_{xy}$, we need to identify an appropriate joint distribution η, that minimizes the total probability for the upper quantile. It turns out that we can formulate this as a linear program shown below (see Figure 2), with marginal and non-negativity constraints. Similar to Algorithm 2, we have Algorithm COMPUTE-VARIABILITY-UPPER-QUANTILE that computes the value $\nu_u(r, \eta_a, \eta_b)$ exactly.

Lemma 27 *Given a feasible median value $r \in \{-(k-1), \cdots, (k-1)\}$ and marginals η_a, η_b, Algorithm 3 computes the value $\nu_u(r, \eta_a, \eta_b)$ exactly and runs in time $O(k)$.*

Proof A similar proof as that of Lemma 26. ■

D.0.1. PROOF OF LEMMA 25

Proof Combining Lemma 26 and Lemma 27, gives us the lemma. ■

$$
\begin{aligned}
\underline{\text{Objective}}: \quad & \min \sum_{x,y} \mathbf{1}\{x - y \le r\}\eta_{xy} \\
& \sum_{y \in \text{PO}_k} \eta_{xy} = \eta_a(x) \quad \forall x \in \text{PO}_k \\
& \sum_{x \in \text{PO}_k} \eta_{xy} = \eta_b(y) \quad \forall y \in \text{PO}_k \\
& \eta_{xy} \ge 0 \quad \forall x, y \in \text{PO}_k \times \text{PO}_k
\end{aligned}
$$

Figure 2: Linear Program for computing the upper quantile component of variability

Algorithm 3 COMPUTE-VARIABILITY-UPPER-QUANTILE

Input: Marginal distributions η_a, η_b and feasible median value $r \in \{-(k-1), \cdots, (k-1)\}$.
Output: $\nu_u(r, \eta_a, \eta_b)$.
$\forall x \in \text{PO}_k, y \in \text{PO}_k$, set $\eta_{xy} \leftarrow 0$
for $y \in \{k-1, k-2, \cdots, 0\}$ **do**
  for $x \in \{y + r, y + r - 1, \cdots, 0\}$ **do**
    Identify all the marginal constraints in LP 2 containing η_{xy} and then slowly increase the value of η_{xy} until one of them becomes tight.
    We mark the corresponding column or row with the tight constraint as *frozen* and do not update any values in them.
  end for
end for
Return $\nu_l(r, \eta_a, \eta_b) \leftarrow \sum_{x,y \in \text{PO}_k \times \text{PO}_k} \mathbf{1}\{x - y \le r\}\eta_{xy}$.