Causal Interaction in Factorial Experiments: Application to Conjoint Analysis*

Naoki Egami† Kosuke Imai‡

December 12, 2016

Abstract

Social scientists use conjoint analysis, which is based on randomized experiments with a factorial design, to analyze multidimensional preferences in a population. In such experiments, several factors, each with multiple levels, are randomized to form a large number of possible treatment conditions. To explore causal interaction in factorial experiments, we propose a new definition of causal interaction effect, called the *average marginal interaction effect* (AMIE). Unlike the conventional interaction effect, the relative magnitude of the AMIE does not depend on the choice of baseline conditions, making its interpretation intuitive even for high-order interaction. We show that the AMIE can be nonparametrically estimated using the ANOVA regression with weighted zero-sum constraints. These two properties enable us to directly regularize the AMIEs by collapsing levels and selecting factors within a penalized ANOVA framework. This reduces false discovery rate and further facilitates interpretation. Finally, we apply the proposed methodology to the conjoint analysis of ethnic voting behavior in Africa and find clear patterns of causal interaction between politicians' ethnicity and their prior records. The proposed methodology is implemented in the open source software.

Key words: ANOVA, causal inference, heterogeneous treatment effects, interaction effects, randomized experiments, regularization

*The proposed methods are implemented through open-source software `FindIt` (Egami *et al.*, 2015), which is freely available as an R package at the Comprehensive R Archive Network (CRAN http://cran.r-project.org/package=FindIt). We thank Elizabeth Carlson for providing us with data and answering our questions. We are also grateful for Jens Hainmueller, Walter Mebane, Dustin Tingley, Teppei Yamamoto, Tyler VanderWeele, and seminar participants at Carnegie Mellon University (Statistics), Georgetown University (School of Public Policy), Stanford (Political Science), Umea University (Statistics), University of Bristol (Mathematics), and UCLA (Political Science) for helpful comments on an earlier version of the paper.

†Ph.D. student, Department of Politics, Princeton University, Princeton NJ 08544. Email: negami@princeton.edu, URL: http://scholar.princeton.edu/negami

‡Professor, Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

1 Introduction

Statistical interaction among treatment variables can be interpreted as causal relationships when the treatments are randomized in an experiment. Causal interaction plays an essential role in the exploration of heterogeneous treatment effects. This paper develops a framework for studying causal interaction in randomized experiments with a factorial design, in which there are multiple factorial treatments with each having several levels. The goal of causal interaction analysis is to identify the combinations of treatments that induce large additional effects beyond the sum of effects separately attributable to each treatment.

Our motivating application is conjoint analysis, which is a type of randomized survey experiment with a factorial design. Conjoint analysis has been extensively used in marketing research (e.g., Green *et al.*, 2001; Marshall and Bradlow, 2002). In a typical application, respondents are asked to evaluate several pairs of randomly selected product profiles defined by multiple factors. The reported preference data are then used to predict consumer demand and sales of potential products.

Recently, conjoint analysis has also gained its popularity among medical and social scientists who study multidimensional preferences among a population of individuals (e.g., Marshall *et al.*, 2010; Hainmueller and Hopkins, 2015). In this paper, we focus on the latter use of conjoint analysis by estimating population average causal effects. Specifically, we analyze a conjoint analysis about coethnic voting in Africa to examine the conditions under which voters prefer political candidates of the same ethnicity (see Section 2 for the details of the experiment and Section 6 for our empirical analysis).

One important limitation of conjoint analysis, as currently conducted in applied research, is that causal interactions are largely ignored. This is unfortunate because studies of multi-dimensional choice necessarily involve the consideration of interaction effects. However, the exploration of causal interactions in conjoint analysis is often difficult for two reasons. First, the relative magnitude of the conventional causal interaction effect depends on the choice of baseline condition. This is problematic because many factors used in conjoint analysis do not have natural baseline conditions (e.g., gender, racial groups, religions, occupations). Second, a typical conjoint analysis has several factors with each having multiple levels. This means that we must apply a regularization method to reduce false discovery and facilitate interpretation. Yet, the lack of invariance to the choice of baseline condition makes the direct application of many regularization methods difficult.

To overcome these problems, we propose an alternative definition of causal interaction effect that is invariant to the choice of baseline condition, making its interpretation intuitive even for high-order interaction (Sections 3 and 4). We call this new causal quantity of interest, the *average marginal interaction effect* (AMIE), because it marginalizes the other treatments rather than conditioning on their baseline values as done in the conventional causal interaction effect. The proposed approach enables researchers to effectively summarize the structure of causal interaction in high-dimension by decomposing the total effect of any treatment combination into the separate effect of each treatment and their interaction effects.

Finally, we also establish the identification condition and develop estimation strategies for the AMIE (Section 5). We propose a nonparametric estimator of the AMIE and show that this estimator be obtained using ANOVA with weighted zero-sum constraints (Scheffe, 1959). Exploiting this equivalence relationship, we apply the method proposed by Post and Bondell (2013) and directly regularize the AMIEs within the ANOVA framework by collapsing levels and selecting factors. Since the AMIE is invariant to the choice of baseline condition, our regularization also has the same invariance property. This also enables a proper regularization of the conditional average effects, which can be computed using the AMIEs. All of our theoretical results

Factors	Levels	
`Coethnicity`	`Yes`	a coethnic of a respondent
	`No`	not a coethnic of a respondent
`Record`	`Yes/Village`	politician for a village with record of good prior record
	`Yes/District`	politician for a district with record of good prior record
	`Yes/MP`	member of parliament with record of good prior record
	`No/Village`	politician for a village with no record of good prior record
	`No/District`	politician for a district with no record of good prior record
	`No/MP`	member of parliament with record of no good prior record
	`No/Business`	businessman with no record of good prior record
`Platform`	`Job`	promise to create new jobs
	`Clinic`	promise to create clinics
	`Education`	promise to improve education
`Degree`	`Yes`	masters degree in business, law, economics, or development
	`No`	bachelors degree in tourism, horticulture, forestry or theater

Table 1: Levels of Four Factors from the Conjoint Analysis in Carlson (2015).

and estimation strategies are shown to hold for causal interaction of any order.

Our paper builds on the causal inference and experimental design literatures that are concerned about interaction effects (see e.g., Cox, 1984; Jaccard and Turrisi, 2003; de González and Cox, 2007; VanderWeele and Knol, 2014). In addition, we draw upon the recent papers that provide the potential outcomes framework for causal inference with factorial experiments and conjoint analysis (Dasgupta *et al.*, 2015; Hainmueller *et al.*, 2014; Lu, 2016a,b). Finally, this paper is also related to the literature on heterogeneous treatment effects, in which different combinations of treatments may exhibit varying degrees of causal effects (e.g., Imai and Ratkovic, 2013; Grimmer *et al.*, 2016). However, much of this literature focus on the interaction between a single treatment and pre-treatment covariates rather than the interaction among multiple treatments (e.g., Hill, 2012; Green and Kern, 2012; Wager and Athey, 2015).

2 Conjoint Analysis of Ethnic Voting

In this paper, we examine a recent conjoint analysis conducted to study coethnic voting in Uganda (Carlson, 2015). Coethnic voting refers to the tendency of some voters to prefer political candidates whose ethnicity is the same as their own. Researchers have observed that coethnic voting occurs frequently among African voters, but the identification of causal effects is often difficult because the ethnicity of candidates are often correlated with other characteristics that may influence voting behavior. To address this problem, the original author conducted a conjoint analysis, in which respondents were asked to choose one of the two hypothetical candidates whose attributes were randomly assigned.

For the experiment, a total of 547 respondents were sampled from villages in Uganda. We analyze a subset of 544 observations after removing 3 observations with missing data. Each respondent was given the description of three pairs of hypothetical presidential candidates. They were then asked to cast a vote for one of the candidates within each pair. These hypothetical candidates are characterized by a total of four factors shown in Table 1: `Coethnicity` (2 levels), `Record` (7 levels), `Platform` (3 levels), and `Degree` (2 levels).

While the levels of all factors are randomly and independently selected for each hypothetical candidate, the distribution of candidate ethnicity depends on the local ethnic diversity so that enough respondents share the same ethnicity as their assigned hypothetical candidates. The original analysis was based on a mixed effects logistic regression with a respondent random effect. While previous studies showed that many voters unconditionally favor coethnic candidates, Carlson (2015) found that voters tend to favor only coethnic candidates with good prior record.

We focus on two methodological challenges of the original analysis. First, the author tests the existence of causal interaction between `Coethnicity` and `Record`, but does not explicitly estimate causal interaction effects. We propose a definition of causal interaction effects in randomized experiments with a factorial design and show how to estimate them. Second, the author dichotomized two factors, `Record` and `Platform`, which have more than two levels and does not have a natural baseline condition. We show how to use a data-driven regularization method when estimating causal interaction effects in a high-dimensional setting. Our reanalysis of this experiment appears in Section 6.

3 Two-Way Causal Interaction

In this section, we introduce a new causal quantity, the *average marginal interaction effect* (AMIE), and show that, unlike the conventional causal interaction effect, it is invariant to the choice of baseline condition. The invariance property enables simple interpretation and effective regularization even when there are many factors. While this section focuses on two-way causal interaction for the sake of simplicity, all definitions and results will be generalized beyond two-way interaction in Section 4.

3.1 The Setup

Consider a simple random sample of n units from the target population $\mathcal{P}$. Let A_i and B_i be two factorial treatment variables of interest for unit i where L_A and L_B be the number of ordered or unordered levels for factors A and B, respectively. We use a_ℓ and b_m to represent levels of the two factors where $\ell = \{0, 1, \cdots, L_A - 1\}$ and $m = \{0, 1, \cdots, L_B - 1\}$. The support of treatment variables A and B, therefore, is given by $\mathcal{A} = \{a_0, a_1, \cdots, a_{L_A-1}\}$ and $\mathcal{B} = \{b_0, b_1, \cdots, b_{L_B-1}\}$, respectively.

We call a combination of factor levels (a_ℓ, b_m) a *treatment combination*. Thus, in the current set-up, the total number of unique treatment combinations is $L_A \times L_B$. Let $Y_i(a_\ell, b_m)$ denote the potential outcome variable of unit i if the unit receives the treatment combination (a_ℓ, b_m). For each unit, only one of the potential outcome variables can be observed, and the realized outcome variable is denoted by $Y_i = \sum_{a_\ell \in \mathcal{A}, b_m \in \mathcal{B}} \mathbf{1}\{A_i = a_\ell, B_i = b_m\} Y_i(a_\ell, b_m)$, where $\mathbf{1}\{A_i = a_\ell, B_i = b_m\}$ is an indicator variable taking 1 when $A_i = a_\ell$ and $B_i = b_m$, and taking 0 otherwise. In this paper, we make the stability assumption, which states that there is neither interference between units nor different versions of the treatment (Cox, 1958; Rubin, 1990).

In addition, we assume that the treatment assignment is randomized.

$$\{Y_i(a_\ell, b_m)\}_{a_\ell \in \mathcal{A}, b_m \in \mathcal{B}} \quad \perp\!\!\!\perp \quad \{A_i, B_i\} \quad \text{for all } i = 1, \ldots, n \tag{1}$$

$$\Pr(A_i = a_\ell, B_i = b_m) \quad > \quad 0 \quad \text{for all } a_\ell \in \mathcal{A} \quad \text{and} \quad b_m \in \mathcal{B} \tag{2}$$

This assumption rules out the use of fractional factorial designs where certain combinations of treatments have zero probability of occurrence. In some cases, however, researchers may wish to eliminate certain treatment combinations for substantive reasons. The standard recommendation is to set the probability for those treatment combinations to small non-zero values under a full factorial design so that the assumption continues to hold (see Hainmueller *et al.*, 2014, footnote 18). Another possibility is to restrict one's analysis to a subset of the data so that the assumption is satisfied.

Under this setup, we review two non-interactive causal effects of interest. First, we define the *average combination effect* (ACE), which represents the average causal effect of a treatment combination $(A_i, B_i) = (a_\ell, b_m)$ relative to a pre-specified baseline condition (a_0, b_0) (e.g., Dasgupta *et al.*, 2015).

$$\tau_{AB}(a_\ell, b_m; a_0, b_0) \quad \equiv \quad \mathbb{E}\{Y_i(a_\ell, b_m) - Y_i(a_0, b_0)\} \tag{3}$$

where $a_\ell, a_0 \in \mathcal{A}$ and $b_m, b_0 \in \mathcal{B}$.

Another causal quantity of interest is the *average marginal effect* (AME). For each unit, we define the marginal effect of treatment condition $A_i = a_\ell$ relative to a baseline condition a_0 by averaging over the distribution of the other treatment B_i. Then, the AME is the population average of this unit-level marginal effect (e.g., Hainmueller *et al.*, 2014; Dasgupta *et al.*, 2015).

$$\psi_A(a_\ell, a_0) \quad \equiv \quad \mathbb{E}\left[\int \{Y_i(a_\ell, B_i) - Y_i(a_0, B_i)\}\, dF(B_i)\right] \tag{4}$$

where $a_\ell, a_0 \in \mathcal{A}$ and B_i is another factor whose distribution function is $F(B_i)$. The AME of b_m relative to b_0, i.e., $\psi_B(b_m, b_0)$, can be defined similarly.

We emphasize that while these two causal quantities require the specification of baseline conditions, the relative magnitude is not sensitive to this choice. For example, if we sort the ACEs by their relative magnitude, the resulting order does not depend on the values of the treatment variables selected for the baseline conditions (a_0, b_0). The same property is applicable to the AMEs where the choice of baseline condition a_0 does not alter their relative magnitude.

3.2 The Average Marginal Interaction Effect

We propose a new definition of two-way causal interaction effect, the *average marginal interaction effect* (AMIE), which is useful for randomized experiments with a factorial design. For each unit, a marginal interaction effect represents the causal effect induced by the treatment combination beyond the sum of the marginal effects separately attributable to each treatment. The AMIE is the population average of this unit-level marginal interaction effect. Specifically, the two-way AMIE of treatment combination (a_ℓ, b_m), with baseline condition (a_0, b_0), is defined as,

$$\begin{aligned} \pi_{AB}(a_\ell, b_m; a_0, b_0) \;&\equiv\; \mathbb{E}\Big[Y_i(a_\ell, b_m) - Y_i(a_0, b_0) - \int \{Y_i(a_\ell, B_i) - Y_i(a_0, B_i)\} dF(B_i) \\ &\qquad\qquad - \int \{Y_i(A_i, b_m) - Y_i(A_i, b_0)\} dF(A_i) \Big] \\ &= \; \tau_{AB}(a_\ell, b_m; a_0, b_0) - \psi_A(a_\ell, a_0) - \psi_B(b_m, b_0) \end{aligned} \tag{5}$$

where $a_\ell, a_0 \in \mathcal{A}$ and $b_m, b_0 \in \mathcal{B}$, $\pi_{AB}(a_\ell, b_m; a_0, b_0)$ is the AMIE, and $\psi(\cdot, \cdot)$ is the AME defined in equation (4).

The AMIE is closely connected to the conventional definition of the *average interaction effect* (AIE). In the causal inference literature (e.g., Cox, 1984; VanderWeele, 2015; Dasgupta *et al.*, 2015), researchers define the AIE of treatment combination (a_ℓ, b_m) relative to baseline condition (a_0, b_0) as,

$$\xi_{AB}(a_\ell, b_m; a_0, b_0) \;\equiv\; \mathbb{E}\{Y_i(a_\ell, b_m) - Y_i(a_0, b_m) - Y_i(a_\ell, b_0) + Y_i(a_0, b_0)\} \tag{6}$$

where $a_\ell, a_0 \in \mathcal{A}$ and $b_m, b_0 \in \mathcal{B}$.

Similar to the AMIE, the AIE has an *interactive effect interpretation*, representing the additional average causal effect induced by the treatment combination beyond the sum of the average causal effects separately attributable to each treatment. This interpretation is based on the following algebraic equality,

$$\xi_{AB}(a_\ell, b_m; a_0, b_0) = \tau_{AB}(a_\ell, b_m; a_0, b_0) - \mathbb{E}\{Y_i(a_\ell, b_0) - Y_i(a_0, b_0)\} - \mathbb{E}\{Y_i(a_0, b_m) - Y_i(a_0, b_0)\}.$$

The difference between the AMIE and the AIE is that the former subtracts the AMEs from the ACE while the latter subtracts the sum of two separate effects due to $A_i = a_\ell$ and $B_i = b_m$ while holding the other treatment variable at its baseline value, i.e., $A_i = a_0$ or $B_i = b_0$.

In addition, the AIE has a *conditional effect interpretation*,

$$\xi_{AB}(a_\ell, b_m; a_0, b_0) = \mathbb{E}\{Y_i(a_\ell, b_m) - Y_i(a_0, b_m)\} - \mathbb{E}\{Y_i(a_\ell, b_0) - Y_i(a_0, b_0)\},$$

which denotes the difference in the average causal effect of $A_i = a_\ell$ relative to $A_i = a_0$ between the two scenarios, one when $B_i = b_m$ and the other when $B_i = b_0$. When such conditional effects are of interest, the AMIE can be used to obtain them. For example, we have,

$$\mathbb{E}\{Y_i(a_\ell, b_0) - Y_i(a_0, b_0)\} = \psi_A(a_\ell; a_0) + \pi_{AB}(a_\ell, b_0; a_0, b_0). \quad (7)$$

Clearly, the scientific question of interest should determine the choice of interpretation. In Section 6, we illustrate how to use the AMIEs for estimating the average conditional effects when necessary.

Finally, the AMIE and the AIE are linear functions of one another. This result is presented below as a special case of Theorem 1 presented in Section 4.

Result 1 (Relationships between the two-way AMIE and two-way AIE) *The two-way average marginal interaction effect (AMIE), defined in equation* (5), *equals the following linear function of the two-way average interaction effects (AIEs), defined in equation* (6).

$$\pi_{AB}(a_\ell, b_m; a_0, b_0) = \xi_{AB}(a_\ell, b_m; a_0, b_0) - \sum_{a \in \mathcal{A}} \Pr(A_i = a)\ \xi_{AB}(a, b_m; a_0, b_0)$$

$$-\sum_{b \in \mathcal{B}} \Pr(B_i = b)\ \xi_{AB}(a_\ell, b; a_0, b_0)$$

Likewise, the AIE can be expressed as the following linear function of the AMIEs.

$$\xi_{AB}(a_\ell, b_m; a_0, b_0) \;=\; \pi_{AB}(a_\ell, b_m; a_0, b_0) - \pi_{AB}(a_\ell, b_0; a_0, b_0) - \pi_{AB}(a_0, b_m; a_0, b_0).$$

Result 1 shows implies that all the AMIEs are zero if and only if all the AIEs are zero. Thus, testing the absence of causal interaction can be done by a F-test, investigating either all the AIEs or all the AMIEs are zero. All causal estimands introduced in this section are identifiable under the assumption of randomized treatment assignment (i.e., equations (1) and (2)).

3.3 Invariance to the Choice of Baseline Condition

One advantage of the AMIE is its invariance to the choice of baseline condition. That is, the relative difference of any pair of AMIEs remains unchanged even if one chooses a different baseline condition. Most causal effects, including the ACE and the AME, have this invariance property. In contrast, the relative magnitude of any two AIEs depends on the choice of baseline condition unless all AIEs are zero. The invariance property is important because without it researchers cannot systematically compare interaction effects of different treatment combinations. Result 2 is a special case of Theorem 2 presented in Section 5.

Result 2 (Invariance to the Choice of Baseline Condition) *The average marginal interaction effect (AMIE), defined in equation* (5)*, is interval invariant whereas the average interaction effect (AIE), defined in equation* (6) *is not. That is, the following statements generally hold,*

$$\pi_{AB}(a_\ell, b_m; a_0, b_0) - \pi_{AB}(a_{\ell'}, b_{m'}; a_0, b_0) = \pi_{AB}(a_\ell, b_m; a_{\tilde{\ell}}, b_{\tilde{m}}) - \pi_{AB}(a_{\ell'}, b_{m'}; a_{\tilde{\ell}}, b_{\tilde{m}})$$
$$\xi_{AB}(a_\ell, b_m; a_0, b_0) - \xi_{AB}(a_{\ell'}, b_{m'}; a_0, b_0) \neq \xi_{AB}(a_\ell, b_m; a_{\tilde{\ell}}, b_{\tilde{m}}) - \xi_{AB}(a_{\ell'}, b_{m'}; a_{\tilde{\ell}}, b_{\tilde{m}})$$

for any $(a_\ell, b_m) \neq (a_{\ell'}, b_{m'})$ *and* $(a_0, b_0) \neq (a_{\tilde{\ell}}, b_{\tilde{m}})$*. In addition, the AIE is interval invariant if and only if all the AIEs are zero. Note that the above differences of the AMIEs are also equal to another AMIE,* $\pi_{AB}(a_\ell, b_m; a_{\ell'}, b_{m'})$*.*

The sensitivity of the AIEs to the choice of baseline condition can be further illustrated by the fact that the AIE of any treatment combination pertaining to one of levels in the baseline condition is equal to zero. That is, if (a_0, b_0) is the baseline condition, then $\xi_{AB}(a_0, b_m; a_0, b_0) = \xi_{AB}(a_\ell, b_0; a_0, b_0) = 0$. If the researchers are only interested in the conditional effect interpretation of the AIEs, these zero AIEs are not of interest. However, this restriction is problematic for the interactive effect interpretation especially when no natural baseline condition exists. In such circumstances, zero AIEs make it impossible to explore all relevant causal interaction effects. To the contrary, researchers need not to restrict their quantities of interest when using the AMIE, which can take a non-zero value even when one treatment is set to the baseline condition. For example, the AMIE can be positive if the effect of the second treatment is large when the first treatment is set to its baseline value.

While it is invariant to the choice of baseline condition, the AMIE critically depends on the distribution of treatments, i.e., $P(A, B)$. This is because the AMIE is a function of the AMEs, which are obtained by marginalizing out other treatments. This dependency of causal quantities is not new. The potential outcomes framework for 2^k factorial experiments introduced by Dasgupta *et al.* (2015), for example, defines causal estimands based on the uniform distribution of treatments.

In contrast, the AMIE is defined using a general treatment distribution. Although the uniform distribution would be a reasonable default choice for many experimentalists, researchers can improve the external validity of their experiment by using a treatment distribution based on the target population (Hainmueller *et al.*, 2014). This is important for the conjoint analysis, in which treatments are often characteristics of people, such as the attributes of politicians in our empirical application (see Section 2). In addition, many researchers already rely on the treatment distribution (Hainmueller *et al.*, 2014). They often independently randomize multiple treatments and then estimate the AME of each treatment by simply ignoring the other treatments. This estimation procedure implicitly conditions on the empirical distribution of treatment assignments.

4 Generalization to Higher Order Interaction

In this section, we generalize the two-way AMIE introduced in Section 3 to higher order causal interaction with more than two factors. We prove that a higher order AMIE retains the same desirable properties and intuitive interpretation.

4.1 The Setup

Suppose that we have a total of J factorial treatments denoted by an vector $\mathbf{T}_i = (T_{i1}, T_{i2}, \cdots, T_{iJ})$ where $J \geq 2$ and each factor T_{ij} has a total of L_j levels. Without loss of generality, let $\mathbf{T}_i^{1:K}$ be a subset of K treatments of interest where $K \leq J$ whereas $\mathbf{T}_i^{(K+1):J}$ denotes the remaining $(J-K)$ factorial treatment variables, which are not of interest. As before, we assume that the treatment assignment is randomized.

Assumption 1 (Randomized Treatment Assignment)

$$Y_i(\mathbf{t}) \quad \perp\!\!\!\perp \quad \mathbf{T}_i \quad \text{and} \quad \Pr(\mathbf{T}_i = \mathbf{t}) \; > \; 0 \quad \text{for all } \mathbf{t}$$

In addition, we assume that J factorial treatments are independent of one another.

Assumption 2 (Independent Treatment Assignment)

$$T_{ij} \quad \perp\!\!\!\perp \quad \mathbf{T}_{i,-j} \quad \text{for all } j \in \{1, 2, \ldots, J\}$$

where $\mathbf{T}_{i,-j}$ *denotes the* $(J-1)$ *factorial treatments excluding* T_{ij}*.*

Assumption 2 is not required for some of the results obtained below, but it considerably simplifies the notation.

We now generalize the definition of the two-way ACE given in equation (3) by accommodating more than two factorial treatments of interest $\mathbf{T}_i^{1:K}$ while allowing for the existence of additional treatments $\mathbf{T}_i^{(K+1):J}$, which are marginalized out.

Definition 1 (The K-way Average Combination Effect) *The K-way average combination effect (ACE) of treatment combination $\mathbf{T}_i^{1:K} = \mathbf{t}^{1:K}$ relative to baseline condition $\mathbf{T}_i^{1:K} = \mathbf{t}_0^{1:K}$ is defined as,*

$$\tau_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) \equiv \mathbb{E}\left[\int \left\{Y_i(\mathbf{T}_i^{1:K} = \mathbf{t}, \mathbf{T}_i^{(K+1):J}) - Y_i(\mathbf{T}_i^{1:K} = \mathbf{t}_0^{1:K}, \mathbf{T}_i^{(K+1):J})\right\} dF(\mathbf{T}_i^{(K+1):J})\right]$$

The generalization of the AME defined in equation (4) to this setting is straightforward. For example, the AME of T_{i1} is obtained by marginalizing the remaining factors $\mathbf{T}_i^{2:J}$ out.

4.2 The K-way Average Marginal Interaction Effect

We now extend the definition of the two-way AMIE, given in equation (5), to higher-order causal interaction and discuss its relationships with the conventional higher-order causal interaction effect. We define the K-way AMIE as the additional effect of treatment combination beyond the sum of all lower-order AMIEs.

Definition 2 (The K-way Average Marginal Interaction Effect) *The K-way average marginal interaction effect (AMIE) of treatment combination $\mathbf{T}_i^{1:K} = \mathbf{t}^{1:K}$, relative to baseline condition, $\mathbf{T}_i^{1:K} = \mathbf{t}_0^{1:K}$, is given by,*

$$\begin{aligned}\pi_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) &\equiv \mathbb{E}\left[\tau_{1:K}^{(i)}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) - \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_k}^{(i)}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k})\right] \\ &= \tau_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) - \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k})\end{aligned}$$

where $\mathcal{K}_k \subseteq \mathcal{K}_K = \{1, \ldots, K\}$ such that $|\mathcal{K}_k| = k$ with $k = 1, \ldots, K$ and $\pi_{1:K}^{(i)}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K})$ is the unit-level K-way marginal interaction effect.

This definition reduces to equation (5) when $K = 2$ because the one-way AMIE is equal to the AME, i.e., $\pi_1(t; t_0) = \psi_1(t, t_0)$.

As in the two-way case, the K-way AMIE is closely related to the K-way AIE. To generalize the two-way AIE given in equation (6), we first define the two-way AIE of treatment combination $\mathbf{t}^{1:2} = (t_1, t_2)$, relative to baseline condition $\mathbf{t}_0^{1:2} = (t_{01}, t_{02})$ by marginalizing the remaining treatments $\mathbf{T}^{3:J}$. The unit-level two-way interaction effect and the two-way AIE are defined as,

$$\xi_{1:2}(\mathbf{t}^{1:2}; \mathbf{t}_0^{1:2}) \equiv \mathbb{E}\left[\int \left\{Y_i(t_1, t_2, \mathbf{T}_i^{3:J}) - Y_i(t_{01}, t_2, \mathbf{T}_i^{3:J}) - Y_i(t_1, t_{02}, \mathbf{T}_i^{3:J}) + Y_i(t_{01}, t_{02}, \mathbf{T}_i^{3:J})\right\} dF(\mathbf{T}_i^{3:J})\right]$$

In addition, define the *conditional* two-way AIE by fixing the level of another treatment T_{i3} at t^*.

$$\begin{aligned}
& \xi_{1:2}(\mathbf{t}^{1:2}; \mathbf{t}_0^{1:2} \mid T_{i3} = t^*) \\
\equiv \quad & \mathbb{E}\left[\int \{Y_i(t_1, t_2, t^*, \mathbf{T}_i^{4:J}) - Y_i(t_{01}, t_2, t^*, \mathbf{T}_i^{4:J}) - Y_i(t_1, t_{02}, t^*, \mathbf{T}_i^{4:J}) + Y_i(t_{01}, t_{02}, t^*, \mathbf{T}_i^{4:J})\} dF(\mathbf{T}_i^{4:J})\right]
\end{aligned}$$

Then, the three-way AIE can be defined as the difference between the ACE of treatment combination $\mathbf{t}^{1:3} = (t_1, t_2, t_3)$ and the sum of all conditional two-way and one-way AIEs while conditioning on the baseline condition $\mathbf{t}_0^{1:3} = (t_{01}, t_{02}, t_{03})$,

$$\begin{aligned}
& \xi_{1:3}(\mathbf{t}^{1:3}; \mathbf{t}_0^{1:3}) \\
= \quad & \tau_{1:3}(\mathbf{t}^{1:3}; \mathbf{t}_0^{1:3}) - \left\{\xi_{1:2}(\mathbf{t}^{1:2}; \mathbf{t}_0^{1:2} \mid T_{i3} = t_{03}) + \xi_{2:3}(\mathbf{t}^{2:3}; \mathbf{t}_0^{2:3} \mid T_{i1} = t_{01}) + \xi_{1,3}(\mathbf{t}^{1,3}; \mathbf{t}_0^{1,3} \mid T_{i2} = t_{02})\right\} \\
& - \left\{\xi_1(t_1; t_{01} \mid \mathbf{T}_i^{2:3} = \mathbf{t}_0^{2:3}) + \xi_2(t_2; t_{02} \mid \mathbf{T}_1^{1,3} = \mathbf{t}_0^{1,3}) + \xi_3(t_3; t_{03} \mid \mathbf{T}_i^{1:2} = \mathbf{t}_0^{1:2})\right\} \qquad (8)
\end{aligned}$$

Note that the one-way conditional AIEs are equivalent to the one-way ACEs or the average effects of single treatments while holding the other treatments at their base level. For example, $\xi_1(t_1; t_{01} \mid T_i^{2:3} = \mathbf{t}_0^{2:3})$ is equal to $\tau_{1:3}(t_1, \mathbf{t}_0^{2:3}; \mathbf{t}_0)$. We also note that $\xi_1(t_1; t_{01}) = \psi_1(t_1; t_{01}) = \pi_1(t_1; t_{01})$ holds. In this way, we can generalize the AIE to higher order causal interaction.

Definition 3 (The K-way Average Interaction Effect) *The K-way average interaction effect (AIE) of treatment combination $\mathbf{T}_i^{1:K} = \mathbf{t}^{1:K} = (t_1, \ldots, t_K)$ relative to baseline condition $\mathbf{T}_i^{1:K} = \mathbf{t}_0^{1:K} = (t_{01}, \ldots, t_{0K})$ is given by,*

$$\begin{aligned}
\xi_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) &= \mathbb{E}\left[\tau_{1:K}^{(i)}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) - \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\mathcal{K}_k}^{(i)}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \mathbf{T}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k})\right] \\
&= \tau_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) - \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \mathbf{T}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k})
\end{aligned}$$

where the second summation is taken over the set of all possible $\mathcal{K}_k \subseteq \mathcal{K}_K = \{1, 2, \ldots, K\}$ such that $|\mathcal{K}_k| = k$, $\tau_{1:K}^{(i)}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K})$ is the unit-level combination effect, and $\xi_{\mathcal{K}_k}^{(i)}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \mathbf{T}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k})$ represents the unit-level interaction effect.

While both estimands have similar interpretations, the K-way AMIE differs from the K-way AIE in important ways. First, the AMIE is expressed as a function of its lower-order effects whereas the AIE is based on the lower-order *conditional* AIEs rather than the lower-order AIEs. This implies that we can decompose the K-way ACE as the sum of the K-way AMIE and all lower-order AMIEs.

$$\tau_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) \quad = \quad \sum_{k=1}^{K} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k}). \tag{9}$$

The decomposition is useful for understanding how interaction effects of various order relate to the overall effect of treatment combination. However, because of conditioning on the baseline value, a similar decomposition is not applicable to the AIEs.

Second, in the experimental design literature, the K-way AIE is often interpreted as a conditional interaction effect (see e.g., Jaccard and Turrisi, 2003; Wu and Hamada, 2011). For example, the three-way AIE of treatment combination $\mathbf{T}_i^{1:3} = \mathbf{t}^{1:3} = (t_1, t_2, t_3)$ relative to baseline condition $\mathbf{T}_i^{1:3} = \mathbf{t}_0^{1:3} = (t_{01}, t_{02}, t_{03})$, given in equation (8), can be rewritten as the difference in the conditional two-way AIEs where the third factorial treatment is either set to t_3 or t_{03},

$$\xi_{1:3}(\mathbf{t}^{1:3}; \mathbf{t}_0^{1:3}) \quad = \quad \xi_{1:2}(\mathbf{t}^{1:2}; \mathbf{t}_0^{1:2} \mid T_{i3} = t_3) - \xi_{1:2}(\mathbf{t}^{1:2}; \mathbf{t}_0^{1:2} \mid T_{i3} = t_{03})$$

Lemma 1 shows that this equivalence relationship can be generalized to the K-way AIE (see Appendix A.1).

Unfortunately, as recognized by others (see e.g., Wu and Hamada, 2011, p. 112), although it is useful when $K = 2$, this conditional interpretation faces difficulty when K is greater than three. For example, the three-way AIE has the conditional effect interpretation, characterizing how the conditional two-way AIE varies as a function of the third factorial treatment. However, according to this interpretation, the two-way AIE, which varies according to the second treatment of interest, itself describes how the main effect of one treatment changes as a function of another treatment. This means that the three-way AIE is the conditional effect of another conditional effect, making it difficult for applied researchers to gain an intuitive understanding.

Finally, as in the two-way case, we can express the K-way AMIE and K-way AIE as linear functions of one another. The next theorem summarizes this result.

THEOREM 1 (RELATIONSHIPS BETWEEN THE K-WAY AMIE AND THE K-WAY AIE) *Under Assumption 2, the K-way average marginal interaction effect (AMIE), given in Definition 2, equals the following linear function of the K-way average interaction effects (AIEs), given in Definition 3. That is, for any $\mathbf{t}^{1:K}$ and $\mathbf{t}_0^{1:K}$, we have.*

$$\pi_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) \ = \xi_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) + \sum_{k=1}^{K-1}(-1)^k \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \int \xi_{\mathcal{K}_k}(\mathbf{T}^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_K \setminus \mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_K}) dF(\mathbf{T}^{\mathcal{K}_k})$$

where $\mathcal{K}_k \subseteq \mathcal{K}_K = \{1, \ldots, K\}$ such that $|\mathcal{K}_k| = k$ with $k = 1, \ldots, K$. Likewise, but without requiring Assumption 2, the K-way AIE can be written as the following linear function of the K-way AMIEs.

$$\xi_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) \ = \ \sum_{k=1}^{K}(-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k})$$

Proof is in Appendix A.2. All causal estimands introduced above are identifiable under Assumption 1. We propose nonparametric unbiased estimators in Section 5.

4.3 Invariance to the Choice of Baseline Condition

As is the case for the two-way AMIE, the K-way AMIE is invariant to the choice of baseline condition. In contrast, the K-way AIEs lack this invariance property. The next theorem generalizes Theorem 2 to the K-way causal interaction.

THEOREM 2 (INVARIANCE AND LACK THEREOF TO THE CHOICE OF BASELINE CONDITION) *The K-way average marginal interaction effect (AMIE), given in Definition 2, is interval invariant whereas the average interaction effect (AIE), given in Definition 3 is not. That is, the following statements generally hold,*

$$\begin{aligned}
\pi_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) \ - \ \pi_{1:K}(\tilde{\mathbf{t}}^{1:K}; \mathbf{t}_0^{1:K}) \ &= \ \pi_{1:K}(\mathbf{t}^{1:K}; \tilde{\mathbf{t}}_0^{1:K}) \ - \ \pi_{1:K}(\tilde{\mathbf{t}}^{1:K}; \tilde{\mathbf{t}}_0^{1:K}) \\
\xi_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) \ - \ \xi_{1:K}(\tilde{\mathbf{t}}^{1:K}; \mathbf{t}_0^{1:K}) \ &\neq \ \xi_{1:K}(\tilde{\mathbf{t}}^{1:K}; \tilde{\mathbf{t}}_0^{1:K}) \ - \ \xi_{1:K}(\tilde{\mathbf{t}}^{1:K}; \tilde{\mathbf{t}}_0^{1:K})
\end{aligned}$$

for any treatment combination $\mathbf{t}^{1:K} \neq \tilde{\mathbf{t}}^{1:K}$ and control condition $\mathbf{t}_0^{1:K} \neq \tilde{\mathbf{t}}_0^{1:K}$.

Proof is in Appendix A.3.

5 Estimation and Regularization

In this section, we show how to estimate the AMIE using the general notation introduced in Section 4. For the sake of simplicity, our discussion focuses on the two-way AMIE but we show that all the results presented here can be generalized to the K-way AMIE. We first introduce the nonparametric estimators based on difference in sample means. We then prove that the AMIE can also be nonparametrically estimated using ANOVA with weighted zero-sum constraints (Scheffe, 1959). While ANOVA is mainly used for a balanced design, our approach is applicable to the unbalanced design as well so long as Assumptions 1 and 2 hold. Finally, we show how to directly regularize the AMIEs by collapsing levels and selecting factors (Post and Bondell, 2013). Because of the invariance property of the AMIEs, this regularization method is also invariant to the choice of baseline condition. The proposed method reduces false discovery and facilitates interpretation when there are many factors.

5.1 Difference-in-means Estimators

In the causal inference literature, the following difference-in-means estimators have been used to nonparametrically estimate the ACE and AME (e.g., Hainmueller *et al.*, 2014; Dasgupta *et al.*, 2015).

$$\hat{\tau}_{jj'}(\ell, m; 0, 0) = \frac{\sum_{i=1}^n Y_i \mathbf{1}\{T_{ij} = \ell, T_{ij'} = m\}}{\sum_{i=1}^n \mathbf{1}\{T_{ij} = \ell, T_{ij'} = m\}} - \frac{\sum_{i=1}^n Y_i \mathbf{1}\{T_{ij} = 0, T_{ij'} = 0\}}{\sum_{i=1}^n \mathbf{1}\{T_{ij} = 0, T_{ij'} = 0\}}$$

$$\hat{\pi}_j(\ell; 0) = \frac{\sum_{i=1}^n Y_i \mathbf{1}\{T_{ij} = \ell\}}{\sum_{i=1}^n \mathbf{1}\{T_{ij} = \ell\}} - \frac{\sum_{i=1}^n Y_i \{T_{ij} = 0\}}{\sum_{i=1}^n \mathbf{1}\{T_{ij} = 0\}}$$

These estimators are unbiased only when the treatment assignment distribution of an experimental study is used to define the AMEs and AMIEs. Then, Definition 2 naturally implies the following nonparametric estimator of the two-way AMIE.

$$\hat{\pi}_{jj'}(\ell, m; 0, 0) = \hat{\tau}_{jj'}(\ell, m; 0, 0) - \hat{\psi}_j(\ell; 0) - \hat{\psi}_{j'}(m; 0)$$

Similarly, the nonparametric estimator of higher-order AMIE can be constructed. It is important to emphasize that these nonparametric estimators do not assume the absence of higher-order interactions (Hainmueller *et al.*, 2014).

5.2 Nonparametric Estimation with ANOVA

Alternatively, the AMIEs can be estimated nonparametrically using ANOVA with weighted zero-sum constraints, which is a convex optimization problem (Scheffe, 1959). For example, the two-way AMIE considered above can be estimated by the saturated ANOVA whose objective function is as follows,

$$\sum_{i=1}^{n}\Bigg(Y_i - \mu - \sum_{j=1}^{J}\sum_{\ell=0}^{L_j-1}\beta_{\ell}^{j}\mathbf{1}\{T_{ij}=\ell\} - \sum_{j=1}^{J-1}\sum_{j'>j}\sum_{\ell=0}^{L_j-1}\sum_{m=0}^{L_{j'}-1}\beta_{\ell m}^{jj'}\mathbf{1}\{T_{ij}=\ell, T_{ij'}=m\}$$
$$-\sum_{k=3}^{J}\sum_{\mathcal{K}_k\subset\mathcal{K}_J}\sum_{\mathbf{t}^{\mathcal{K}_k}}\beta_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k}\mathbf{1}\{\mathbf{T}_i^{\mathcal{K}_k}=\mathbf{t}^{\mathcal{K}_k}\}\Bigg)^2 \quad (10)$$

where μ is the global mean, β_{ℓ}^{j} is the coefficient for the first-order term for the jth factor with ℓ level, $\beta_{\ell m}^{jj'}$ is the coefficient for the second-order interaction term for the jth and j'th factors with ℓ and m levels, respectively, and more generally $\beta_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k}$ is the coefficient for the kth interaction term for a set of k factors $\mathcal{K}_k$ when their levels equal to $\mathbf{t}^{\mathcal{K}_k}$. Note that as in Section 4, we have $|\mathcal{K}_k| = k$ and $\mathcal{K}_J = \{1, 2, \ldots, J\}$. We emphasize that the nonparametric estimation requires all interaction terms up to J-way interaction. See Section 5.3 for efficient parametric estimation.

We minimize the objective function given in equation (10) subject to the following weighted zero-sum constraints where the weights are given by the marginal distribution of treatment assignment,

$$\sum_{\ell=0}^{L_j-1}\Pr(T_{ij}=\ell)\beta_{\ell}^{j} \;=\; 0 \quad \text{for all } j, \quad (11)$$
$$\sum_{\ell=0}^{L_j-1}\Pr(T_{ij}=\ell)\beta_{\ell m}^{jj'} \;=\; 0 \quad \text{for all } j\neq j' \text{ and } m\in\{0,1,\ldots,L_{j'}-1\}, \quad (12)$$
$$\sum_{\ell=0}^{L_j-1}\Pr(T_{ij}=\ell)\mathbf{1}\{t_j=\ell\}\beta_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k} \;=\; 0 \quad \text{for all } j,\ \mathbf{t}^{\mathcal{K}_k}, \text{ and } \mathcal{K}_k\subset\mathcal{K}_J \text{ such that } k\geq 3 \text{ and } j\in\mathcal{K}_k$$

(13)

Finally, the next theorem shows that the difference in the estimated ANOVA coefficients represents a nonparametric estimate of the AMIE.

Theorem 3 (Nonparametric Estimation with ANOVA) *Under Assumptions 1 and 2, differences in the estimated coefficients from ANOVA based on equations* (10)–(13) *represent nonparametric unbiased estimators of the AME and the AMIE:*

$$\mathbb{E}(\hat{\beta}^j_\ell - \hat{\beta}^j_0) = \psi_j(\ell; 0), \quad \mathbb{E}(\hat{\beta}^{jj'}_{\ell m} - \hat{\beta}^{jj'}_{00}) = \pi_{jj'}(\ell, m; 0, 0), \quad \mathbb{E}(\hat{\beta}^{\mathcal{K}_k}_{\mathbf{t}^{\mathcal{K}_k}} - \hat{\beta}^{\mathcal{K}_k}_{\mathbf{t}_0^{\mathcal{K}_k}}) = \pi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k})$$

Proof is given in Appendix A.4. These estimators are asymptotically equivalent to their corresponding difference-in-means estimators when the treatment assignment distribution of an experimental study is used as weights. The proposed ANOVA framework, however, allows researchers to use any treatment assignment distributions to define the AME and the AMIE so long as Assumptions 1 and 2 hold.

5.3 Regularization via GASH-ANOVA

A key advantage of this ANOVA-based estimator in Section 5.2 over the difference-in-means estimator in Section 5.1 is that we can directly regularize the AMIEs in a penalized regression framework. The regularization is especially useful for reducing false positives and facilitating interpretation when the number of factors is large.

We apply the regularization method (Grouping and Selection using Heredity in ANOVA or GASH-ANOVA) proposed by Post and Bondell (2013), which places penalties on difference in coefficients of the ANOVA regression. As shown above, these differences correspond to the AMEs and AMIEs. While there exist other regularization methods for categorical variables (e.g., Yuan and Lin, 2006; Meier *et al.*, 2008; Lim and Hastie, 2015; Zhao *et al.*, 2009; Huang *et al.*, 2009, 2012), these methods regularize coefficients rather than their differences. In addition, GASH-ANOVA collapses levels and selects factors by jointly considering the AMEs and AMIEs rather than the AMEs alone. This is attractive because many social scientists believe large interaction effects can exist even when marginal effects are small. The method also collapses levels in a mutually consistent manner.

Finally, because the AMEs and AMIEs are invariant to the choice of baseline condition, this regularization method also inherits the invariance property, which is not generally the case (Lim and Hastie, 2015). In particular, even if one is interested in conditional average causal effects, regularization should be based on the AMEs and AMIEs because of their invariance property. As shown in equation (7), we can compute the conditional average effects directly from these quantities.

To illustrate the application of GASH-ANOVA, consider a situation of practical interest in which we assume the absence of causal interaction higher than the second order. That is, in equation (10), we assume $\beta_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k} = 0$ for all $k \geq 3$. GASH-ANOVA collapses two levels within a factor by directly and jointly regularizing the AMEs and AMIEs that involve those two levels. Define the set of all the AMEs and AMIEs that involve levels ℓ and ℓ' of the jth factor as follows,

$$\phi^j(\ell, \ell') \;\; = \;\; \left\{|\beta_\ell^j - \beta_{\ell'}^j|\right\} \bigcup \left\{\bigcup_{j' \neq j} \bigcup_{m=0}^{L_{j'}-1} |\beta_{\ell m}^{jj'} - \beta_{\ell' m}^{jj'}|\right\}$$

Finally, the penalty is given by,

$$\sum_{j=1}^{J} \sum_{\ell, \ell'} w_{\ell\ell'}^j \max\{\phi^j(\ell, \ell')\} \leq c$$

where c is the cost parameter and $w_{\ell\ell'}^j$ is the adaptive weight of the following form,

$$w_{\ell\ell'}^j \;\; = \;\; \left[(L_j + 1)\sqrt{L_j} \max\{\bar{\phi}^j(\ell, \ell')\}\right]^{-1}$$

where $(L_j + 1)\sqrt{L_j}$ is the standardization factor (Bondell and Reich, 2009), and $\bar{\phi}^j(\ell, \ell')$ represents the corresponding set of all AMEs and AMIEs estimated without regularization. Post and Bondell (2013) show that, when combined with equations (10)–(13), the resulting optimization problem is a quadratic programming problem. They also prove that the method has the oracle property.

6 Empirical Analysis

We apply the proposed method to the conjoint analysis of coethnic voting described in Section 2. Our analysis finds clear patterns of causal interaction between the `Record` and `Coethnicity` variables as well as between the `Record` and `Platform` variables.

6.1 A Statistical Model of Preference Differentials

Our empirical application is based on the choice-based conjoint analysis, in which respondents are asked to evaluate three pairs of hypothetical presidential candidates in turn. Let $Y_i(\mathbf{t})$ be the potential preference by respondent i for a hypothetical candidate characterized by a vector of attributes $\mathbf{t}$. In this experiment, $\mathbf{t}$ is a four dimensional vector, based on the values of factorial treatments shown in Table 1 where each factor T_{ij} has L_j levels (i.e., {`Coethnicity, Record, Platform, Degree`}).

Given the limited sample size, we assume the absence of three-way or higher-order causal interaction and use the following ANOVA regression model of potential outcomes with all one-way effects and two-way interactions.

$$Y_i(\mathbf{t}) \;=\; \mu + \sum_{j=1}^{4} \sum_{\ell=0}^{L_j-1} \beta_\ell^j \mathbf{1}\{t_{ij} = \ell\} + \sum_{j=1}^{4} \sum_{j' \neq j} \sum_{\ell=0}^{L_j-1} \sum_{m=0}^{L_{j'}-1} \beta_{\ell m}^{jj'} \mathbf{1}\{t_{ij} = \ell, t_{ij'} = m\} + \epsilon_i(\mathbf{t}) \tag{14}$$

The results in Section 5.2 implies that the coefficients in this model represent the AIEs and AMIEs.

In this conjoint analysis, respondents evaluate a pair of hypothetical candidates with different attributes. This means that we only observe whether respondent i prefers a candidate with attributes $\mathbf{T}_i^*$ over another candidate with attributes $\mathbf{T}_i^\dagger$. Thus, based on the model of preference given in equation (14), we construct a linear probability model of preference differential,

$$\Pr(Y_i(\mathbf{T}_i^*) > Y_i(\mathbf{T}_i^\dagger) \mid \mathbf{T}_i^*, \mathbf{T}_i^\dagger) \;=\; \tilde{\mu} + \sum_{j=1}^{4} \sum_{\ell=0}^{L_j-1} \beta_\ell^j (\mathbf{1}\{T_{ij}^* = \ell\} - \mathbf{1}\{T_{ij}^\dagger = \ell\})$$

$$+\sum_{j=1}^{4}\sum_{j'\neq j}\sum_{\ell=0}^{L_j-1}\sum_{m=0}^{L_{j'}-1}\beta_{\ell m}^{jj'}(\mathbf{1}\{T_{ij}^*=\ell,T_{ij'}^\dagger=m\}-\mathbf{1}\{T_{ij}^*=\ell,T_{ij'}^\dagger=m\})$$

where $\tilde{\mu}=0.5$ if a position within a pair does not matter. Note that the independence of irrelevant alternatives is assumed. If we additionally assume the difference in errors follow independent Type I extreme value distributions, the model becomes the conditional logit model, which is popular in conjoint analysis (McFadden, 1974).

We minimize the sum of squared residuals, subject to the constraints given in equations (11) and (12) where $\Pr(T_{ij}=\ell)$ represents the marginal distribution of T_{ij}^* and $T_{ij}^\dagger$ together. We also apply the regularization method discussed in Section 5.3. To be consistent with the original dummy coding, we treat `Record` and `Platform` as ordered categorical variables and place penalties on the differences between adjacent levels rather than the differences based on every pairwise comparison. We use the order of levels as shown in Table 1. We choose the uniform distribution for treatment assignment and select the value of the cost parameter c based on the minimum mean squared error criterion in 10-fold cross validation. Since the inference for a regularization method that collapses levels of factorial variables is not established in the literature (Bühlmann and Dezeure, 2016), we focus on the stability of selection (e.g., Breiman, 1996; Meinshausen and Bühlmann, 2010). In particular, we estimate the selection probability for each AME and AMIE using one minus the proportion of 500 bootstrap replicates in which all coefficients for the corresponding factor or factor interaction are estimated to be zero (Efron, 2014; Hastie *et al.*, 2015). Although we do not control the family wise error rate, we follow Meinshausen and Bühlmann (2010) and use 90% cutoff as our default.

6.2 Findings

We begin by reporting the ranges of the estimated AMEs and AMIEs and their selection probability to determine significant factors and factor interactions, respectively.

	Range	Selection prob.
AME		
`Record`	0.122	1.00
`Coethnicity`	0.053	1.00
`Platform`	0.023	0.93
`Degree`	0.000	0.33
AMIE		
`Coethnicity` × `Record`	0.053	1.00
`Record` × `Platform`	0.030	0.92
`Platform` × `Coethnic`	0.008	0.64
`Coethnicity` × `Degree`	0.000	0.62
`Platform` × `Degree`	0.000	0.35
`Record` × `Degree`	0.000	0.09

Table 2: Ranges of the Estimated Average Marginal Effects (AMEs) and Estimated Average Marginal Interaction Effects (AMIEs). The estimated selection probability of the AME (AMIE) is one minus the proportion of 500 bootstrap replicates in which all coefficients for the corresponding factor (factor interaction) are estimated to be zero.

As shown in Table 2, three factors — `Record`, `Platform`, and `Coethnicity` — are found to be significant factors whereas `Degree` is not. In terms of the AMIEs, the interaction `Coethnicity` × `Record`, which is the basis of the main finding in the original article, is estimated to have a large range of 5.3 percentage point, and is selected with probability one. The range of this AMIE is as great as that of the AME of `Coethnicity` and is greater than that of `Platform`. Additionally, the proposed method selects the causal interaction, `Record` × `Platform`, with probability 0.92.

Next, we examine the estimated AMEs presented in Table 3. For the `Record` variable, under the 90% selection probability rule, we collapse a total of original seven levels into three levels – {`Yes/Village`, `Yes/District`, `Yes/MP`}, {`No/Village`, `No/District` , `No/MP`}, and { `No/Businessman`}. This partition suggests that politicians with good record are preferred over those without it including businessman. Similarly, we find two groups in the `Platform` variable – {`Jobs, Clinic`} and {`Education`} – where voters appear to favor candidates with the education platform on average.

We now investigate two significant causal interactions, `Coethnicity` × `Record`

Factor	AME	Selection prob.
Record		
Yes/Village	0.122	⟩ 0.71
Yes/District	0.122	⟩ 0.77
Yes/MP	0.101	⟩ 1.00
No/Village	0.047	⟩ 0.74
No/District	0.051	⟩ 0.74
No/MP	0.047	⟩ 1.00
No/Businessman	base	
Platform		
Jobs	−0.023	⟩ 0.56
Clinic	−0.023	⟩ 0.94
Education	base	
Coethnicity	0.053	1.00
Degree	0.000	0.33

Table 3: The Estimated Average Marginal Effects (AMEs). The estimated selection probability is the proportion of 500 bootstrap replicates in which the difference between two adjacent levels is estimated to be different from zero.

and `Record` × `Platform`. Figure 1 visualizes all estimated AMIEs within each factor interaction. The cells with warmer red (colder blue) color represents a greater (smaller) AMIE than the average AMIE within that factor interaction. The estimates with regularization (right column) show clearer patterns for causal interaction than those without regularization (left column).

First, regarding the `Coethnicity` × `Record` interaction (upper panel of the figure), for example, we find that being coethnic gives an average bonus of 5.3 percentage point if a candidate is an MP with good record beyond the average effect of coethnicity (selec. prob. = 1). In contrast, being coethnic has an additional penalty of 4.6 percentage points when a candidate is a district level politician without good record (selec. prob. = 0.98). As shown in equation (7), we can compute the average conditional effect as the sum of the AME and AMIE. As expected, while the conditional average effect of being coethnic for an MP candidate with good record is 10.7 percentage point (selec. prob. = 1), this effect is almost zero for an MP candidate without good record. These findings support the argument of Carlson (2015).

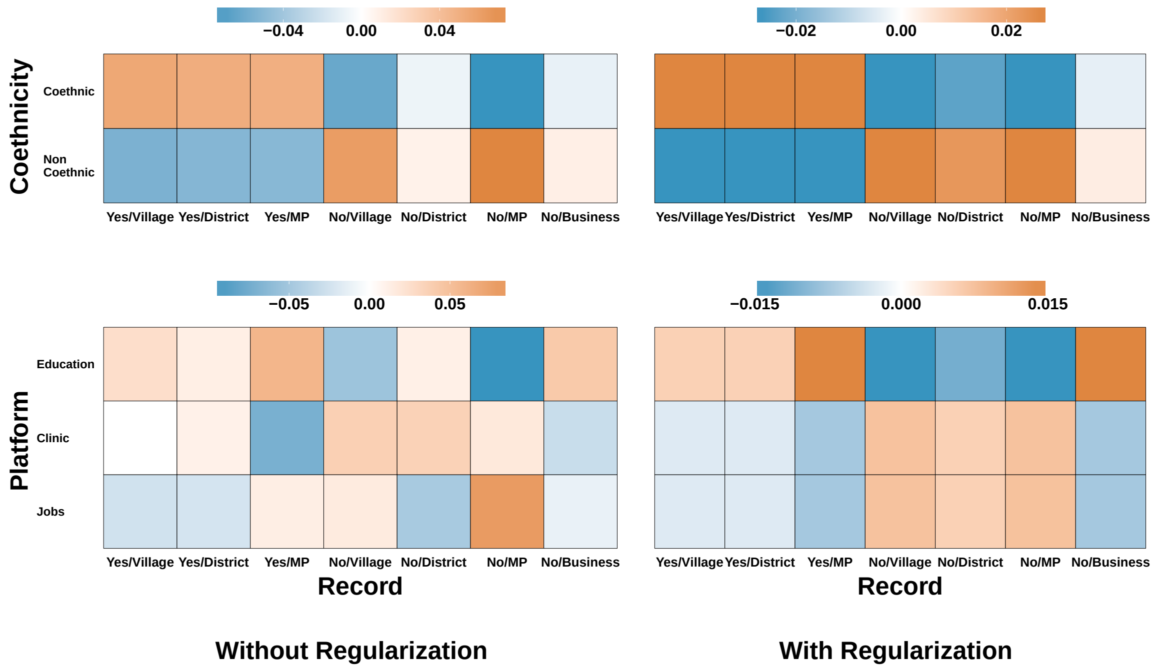

Figure 1: The Estimated AMIEs for `Coethnicity` × `Record` (the first row) and `Platform` × `Record` (the second row). The first and second columns show the estimated AMIEs without and with regularization, respectively.

The decomposition shown in equation (9) can be used to understand the ACE. As an illustration, we decompose the ACE of {`Coethnic, No/Business`} relative to {`Non-coethnic, No/MP`}, which is a estimated negative effect of 2.4 percentage points (selec. prob. = 0.89), as follows,

$$
\begin{aligned}
&\underbrace{\tau(\texttt{Coethnic, No/Business}; \texttt{Non-coethnic, No/MP})}_{-2.4} \\
= \; &\underbrace{\psi(\texttt{Coethnic}; \texttt{Non-coethnic})}_{5.3} + \underbrace{\psi(\texttt{No/Business}; \texttt{No/MP})}_{-4.7} \\
&\qquad + \underbrace{\pi(\texttt{Coethnic, No/Business}; \texttt{Non-coethnic, No/MP})}_{-3.0}
\end{aligned}
$$

We observe that while the average effect of being coethnic is 5.3 percentage points, being a businessman, relative to being an MP without good record, yields an average effect of negative 4.7 percentage points. In addition, being a coethnic businessman has an additional penalty of 3 percentage points relative to non-coethnic MP without good record. All three estimates are selected with probability one.

Finally, we examine the `Platform` $\times$ `Record` interaction, which was not discussed in the original study. We find two distinct groups: (1) politicians with record, businessmen without record and (2) politicians without record. Candidates in the second group appear to receive an additional penalty by promising to improve education. Specifically, the estimated AMIE of {`Education, No/MP`} relative to {`Job, No/MP`} is -2.3 percentage point (selec. prob. = 0.98). In fact, the average conditional effect of `Education` relative to `Job` given `No/MP` is about zero (selec. prob. = 0.66). These results suggest that even though promising to improve education is effective on average (the estimated AME of `Education` relative to `Job` is 2.3 percentage point (selec. prob. = 0.93), it has no effect for politicians without record.

7 Concluding Remarks

In this paper, we propose a new causal interaction effect for randomized experiments with a factorial design, in which there exist many factors with each having several levels. We call this quantity, the average marginal interaction effect (AMIE). Unlike the conventional causal interaction effect, the AMIE is invariant to the choice of baseline. This enables us to provide a simpler interpretation even in a high-dimensional setting. We show how to nonparametrically estimate the AMIE within the ANOVA regression framework. The invariance property also enables us to apply a reguralization method by directly penalizing the AMIEs. This reduces false discovery and facilitates interpretation.

Our method is motivated by and applied to conjoint analysis, a popular survey experiment with a factorial design. The methodological literature on conjoint analysis has largely ignored the role of causal interaction. The method proposed in this paper allows researchers to effectively explore significant causal interaction among several factors. Although not investigated in this paper, future research should investigate interaction between treatments and pre-treatment covariates. It is also of interest to develop sequential experimental designs in the context of factorial experiments so that researchers can efficiently reduce the number of treatments.

References

Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in anova. *Biometrics* **65**, 1, 169–177.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 6, 2350–2383.

Bühlmann, P. and Dezeure, R. (2016). Discussion of 'regularized regression for categorical data' by tutz and gertheiss. *Statistical Modelling* **16**, 3, 205–211.

Carlson, E. (2015). Ethnic voting and accountability in africa: A choice experiment in uganda. *World Politics* **67**, 02, 353–385.

Cox, D. R. (1958). *Planning of Experiments.* John Wiley & Sons, New York.

Cox, D. R. (1984). Interaction. *International Statistical Review* **52**, 1, 1–24.

Dasgupta, T., Pillai, N. S., and Rubin, D. B. (2015). Causal inference from 2^k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **77**, 4, 727–753.

de González, A. B. and Cox, D. R. (2007). Interpretation of interaction: A review. *The Annals of Applied Statistics* **1**, 2, 371–385.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association* **109**, 507, 991–1007.

Egami, N., Ratkovic, M., and Imai, K. (2015). FindIt: Finding heterogeneous treatment effects. available at the Comprehensive R Archive Network (CRAN). `http://CRAN.R-project.org/package=FindIt`.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly* **76**, 3, 491–511.

Green, P. E., Krieger, A. M., and Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces* **31**, 3_supplement, 56–73.

Grimmer, J., Messing, S., and Westwood, S. J. (2016). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Unpublished manuscript, Stanford University, Stanford, CA* .

Hainmueller, J. and Hopkins, D. J. (2015). The hidden american immigration consensus: A conjoint analysis of attitudes toward immigrants. *American Journal of Political Science* **59**, 3, 529–548.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis* **22**, 1, 1–30.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations.* CRC Press.

Hill, J. L. (2012). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 1, 217–240.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science* **27**, 4, 481–499.

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 2, 339–355.

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

Jaccard, J. and Turrisi, R. (2003). *Interaction effects in multiple regression.* Sage Publications.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics* **24**, 3, 627–654.

Lu, J. (2016a). Covariate adjustment in randomization-based causal inference for 2k factorial designs. *Statistics & Probability Letters* **119**, 11–20.

Lu, J. (2016b). On randomization-based and regression-based inferences for 2k factorial designs. *Statistics & Probability Letters* **112**, 72–78.

Marshall, D., Bridges, J. F., Hauber, B., Cameron, R., Donnalley, L., Fyie, K., and Johnson, F. R. (2010). Conjoint analysis applications in health: How are studies being designed and reported? *The Patient: Patient-Centered Outcomes Research* **3**, 4, 249–256.

Marshall, P. and Bradlow, E. T. (2002). A unified approach to conjoint analysis models. *Journal of the American Statistical Association* **97**, 459, 674–682.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka, ed., *Frontiers in econometrics.* Academic Press.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 1, 53–71.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 4, 417–473.

Post, J. B. and Bondell, H. D. (2013). Factor selection and structural identification in the interaction anova model. *Biometrics* **69**, 1, 70–79.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 472–480.

Scheffe, H. (1959). *The analysis of variance*. John Wiley & Sons.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J. and Knol, M. J. (2014). A tutorial on interaction. *Epidemiologic Methods Epidemiol. Methods* **3**, 1, 33–72.

Wager, S. and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342* .

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, vol. 552. John Wiley & Sons.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 1, 49–67.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37**, 6A, 3468–3497.

A Mathematical Appendix: Proofs of Theorems

A.1 Lemmas

Below, we describe all the lemmas, which are used to prove the main theorems of this paper. For completeness, their proofs appear in the supplementary appendix.

Lemma 1 (An Alternative Definition of the K-way Average Interaction Effect) *The K-way average interaction effect (AIE) of treatment combination $\mathbf{T}_i^{1:K} = \mathbf{t}^{1:K} = (t_1, \ldots, t_K)$ relative to baseline condition $\mathbf{T}_i^{1:K} = \mathbf{t}_0^{1:K} = (t_{01}, \ldots, t_{0K})$, given in Definition 3, can be rewritten as,*

$$\xi_{1:K}(\mathbf{t}^{1:K}; \mathbf{t}_0^{1:K}) \;=\; \xi_{1:(K-1)}(\mathbf{t}^{1:(K-1)}; \mathbf{t}_0^{1:(K-1)} \mid T_{iK} = t_K) - \xi_{1:(K-1)}(\mathbf{t}^{1:(K-1)}; \mathbf{t}_0^{1:(K-1)} \mid T_{iK} = t_{0K})$$

Lemma 2 *Under Assumption 2, for any $k = 1, \ldots, K$, the following equality holds,*

$$\int_{\overline{\mathcal{F}}^{\mathcal{K}_k}} \xi_{\mathcal{K}_K}(\tilde{\mathbf{T}}^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_K \setminus \mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_K}) dF(\tilde{\mathbf{T}}^{\mathcal{K}_k}) \;=\; \xi_{\mathcal{K}_K \setminus \mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K \setminus \mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k})$$
$$+ \sum_{\ell=1}^{k} (-1)^{\ell} \sum_{\mathcal{K}_\ell \subseteq \mathcal{K}_k} \int_{\overline{\mathcal{F}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}} \xi_{\mathcal{K}_K \setminus \mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K \setminus \mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k} \mid \tilde{\mathbf{T}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell}) dF(\tilde{\mathbf{T}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell})$$

Lemma 3 (Decomposition of the K-way AIE) *The K-way Average Treatment Interaction Effect (AIE) (Definition 3), can be decomposed into the sum of the K-way conditional Average Treatment Combination Effects (ACEs). Formally, let $\mathcal{K}_k \subseteq \mathcal{K}_K = \{1, \ldots, K\}$ with $|\mathcal{K}_k| = k$ where $k = 1, \ldots, K$. Then, the K-way AIE can be written as follows,*

$$\xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K}) \;=\; \sum_{k=1}^{K} (-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k})$$

where the second summation is taken over the set of all possible $\mathcal{K}_k$ and the k-way conditional ACE is defined as,

$$\tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \;=\; \mathbb{E}\left\{\int_{\underline{\mathcal{F}}^{\mathcal{K}_K}} \{Y_i(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}, \underline{T_i}^{\mathcal{K}_K}) - Y_i(\mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}, \underline{T_i}^{\mathcal{K}_K})\} dF(\underline{\mathbf{T}}_i^{\mathcal{K}_K})\right\}$$

Lemma 4 (Decomposition of the K-way AMIE) *The K-way Average Marginal Treatment Interaction Effect (AMIE), defined in Definition 2, can be decomposed into the sum of the K-way Average Treatment Combination Effects (ACEs). Formally, let*

$\mathcal{K}_k \subseteq \mathcal{K}_K = \{1, \ldots, K\}$ with $|\mathcal{K}_k| = k$ where $k = 1, \ldots, K$. Then, the K-way AMIE can be written as follows,

$$\pi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K}) \;=\; \sum_{k=1}^{K} (-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k}),$$

where the second summation is taken over the set of all possible $\mathcal{K}_k$.

A.2 Proof of Theorem 1

We use proof by induction. Under Assumption 2, we first show for $K = 2$. To simplify the notation, we do not write out the $J-2$ factors that we marginalize out. We begin by decomposing the AME as follows,

$$\begin{aligned}
\psi_A(a_l, a_0) \;&=\; \int_{\mathcal{B}} \mathbb{E}\{Y_i(a_\ell, B_i) - Y_i(a_0, B_i)\}\; dF(B_i) \\
&= \mathbb{E}\{Y_i(a_\ell, b_0) - Y_i(a_0, b_0)\} + \int_{\mathcal{B}} \mathbb{E}\{Y_i(a_\ell, B_i) - Y_i(a_0, B_i) - Y_i(a_\ell, b_0) + Y_i(a_0, b_0)\}\; dF(B_i) \\
&= \mathbb{E}\{Y_i(a_\ell, b_0) - Y_i(a_0, b_0)\} + \int_{\mathcal{B}} \xi_{AB}(a_\ell, B_i; a_0, b_0)\; dF(B_i).
\end{aligned}$$

Similarly, we have $\psi_B(b_m, b_0) = \mathbb{E}\{Y_i(a_0, b_m) - Y_i(a_0, b_0)\} + \int_{\mathcal{A}} \xi_{AB}(A_i, b_m; a_0, b_0)\; dF(A_i)$. Given the definition of the AMIE in equation (5), we have,

$$\begin{aligned}
\pi_{AB}(a_\ell, b_m, a_0, b_0) \;&=\; \mathbb{E}\{Y_i(a_\ell, b_m) - Y_i(a_0, b_0)\} - \psi_A(a_\ell, a_0) - \psi_B(b_m, b_0) \\
&= \xi_{AB}(a_\ell, b_m; a_0, b_0) - \int_{\mathcal{B}} \xi_{AB}(a_\ell, B_i; a_0, b_0)\; dF(B_i) - \int_{\mathcal{A}} \xi_{AB}(A_i, b_m; a_0, b_0)\; dF(A_i)
\end{aligned}$$

This proves that the AMIE is a linear function of the AIEs. We next show that the AIE is also a linear function of the AMIEs.

$$\begin{aligned}
\xi_{AB}(a_\ell, b_m; a_0, b_0) \;&=\; \mathbb{E}[Y_i(a_\ell, b_m) - Y_i(a_0, b_0)] - \psi_A(a_\ell, a_0) - \psi_A(b_m, b_0) \\
&\quad - \mathbb{E}[Y_i(a_\ell, b_0) - Y_i(a_0, b_0)] + \psi_A(a_\ell, a_0) - \mathbb{E}[Y_i(a_0, b_m) - Y_i(a_0, b_0)] + \psi_A(b_m, b_0) \\
&= \pi_{AB}(a_\ell, b_m; a_0, b_0) - \pi_{AB}(a_\ell, b_0; a_0, b_0) - \pi_{AB}(a_0, b_m; a_0, b_0)
\end{aligned}$$

Thus, we obtain the desired results for $K = 2$.

Now we show that if the theorem holds for any K with $K \geq 2$, it also holds for $K+1$. First, using Lemma 2, we rewrite the equation of interest as follows,

$$\pi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K}) = \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K}) + \sum_{k=1}^{K-1}(-1)^k \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \Big\{ \xi_{\mathcal{K}_K \setminus \mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K \setminus \mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k})$$
$$+ \sum_{\ell=1}^{k}(-1)^\ell \sum_{\mathcal{K}_\ell \subseteq \mathcal{K}_k} \int_{\overline{\mathcal{F}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}} \xi_{\mathcal{K}_K \setminus \mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K \setminus \mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k} \mid \tilde{\mathbf{T}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell}) dF(\tilde{\mathbf{T}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}) \Big\}$$

Utilizing the the definition of the K-way AMIE given in Definition 2 and the assumption that the theorem holds for K, we have,

$$\pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) = \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) - \sum_{k=1}^{K} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_{K+1}} \pi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k}),$$
$$= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}})$$
$$- \sum_{k=1}^{K} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_{K+1}} \Bigg[\xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k}) + \sum_{m=1}^{k-1}(-1)^m \sum_{\mathcal{K}_m \subseteq \mathcal{K}_k} \Big\{ \xi_{\mathcal{K}_k \setminus \mathcal{K}_m}(\mathbf{t}^{\mathcal{K}_k \setminus \mathcal{K}_m}, \mathbf{t}_0^{\mathcal{K}_k \setminus \mathcal{K}_m})$$
$$+ \sum_{\ell=1}^{m}(-1)^\ell \sum_{\mathcal{K}_\ell \subseteq \mathcal{K}_m} \int_{\overline{\mathcal{F}}^{\mathcal{K}_m \setminus \mathcal{K}_\ell}} \xi_{\mathcal{K}_k \setminus \mathcal{K}_m}(\mathbf{t}^{\mathcal{K}_k \setminus \mathcal{K}_m}, \mathbf{t}_0^{\mathcal{K}_k \setminus \mathcal{K}_m} \mid \tilde{\mathbf{T}}^{\mathcal{K}_m \setminus \mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell}) dF(\tilde{\mathbf{T}}^{\mathcal{K}_m \setminus \mathcal{K}_\ell}) \Big\} \Bigg], \tag{15}$$

After rearranging equation (15), the coefficient for $\xi_{\mathcal{K}_{K+1} \setminus \mathcal{K}_u}(\mathbf{t}^{\mathcal{K}_{K+1} \setminus \mathcal{K}_u}, \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_u})$ is equal to $(-1)^u$. Similarly, the coefficient of the following term is equal to $(-1)^{u+v}$.

$$\int_{\overline{\mathcal{F}}^{\mathcal{K}_u \setminus \mathcal{K}_v}} \xi_{\mathcal{K}_{K+1} \setminus \mathcal{K}_u}(\mathbf{t}^{\mathcal{K}_{K+1} \setminus \mathcal{K}_u}, \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_u} \mid \tilde{\mathbf{T}}^{\mathcal{K}_u \setminus \mathcal{K}_v}, \overline{\mathbf{T}}_i^{\mathcal{K}_v} = \mathbf{t}_0^{\mathcal{K}_v}) dF(\tilde{\mathbf{T}}^{\mathcal{K}_u \setminus \mathcal{K}_v})$$

Therefore, we can rewrite equation (15) as follows,

$$\pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}})$$
$$= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) + \sum_{k=1}^{K}(-1)^k \sum_{\mathcal{K}_k \subseteq \mathcal{K}_{K+1}} \Big[\xi_{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k})$$
$$+ \sum_{\ell=1}^{k-1}(-1)^\ell \sum_{\mathcal{K}_\ell \subseteq \mathcal{K}_k} \int_{\overline{\mathcal{F}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}} \xi_{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k} \mid \tilde{\mathbf{T}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell}) dF(\tilde{\mathbf{T}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}) \Big]$$
$$= \xi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) + \sum_{k=1}^{K}(-1)^k \sum_{\mathcal{K}_k \subseteq \mathcal{K}_{K+1}} \Big[\xi_{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k})$$
$$+ \sum_{\ell=1}^{k}(-1)^\ell \sum_{\mathcal{K}_\ell \subseteq \mathcal{K}_k} \int_{\overline{\mathcal{F}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}} \xi_{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k} \mid \tilde{\mathbf{T}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell}) dF(\tilde{\mathbf{T}}^{\mathcal{K}_k \setminus \mathcal{K}_\ell}) \Big]$$

$$= \xi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}}) + \sum_{k=1}^{K}(-1)^k \sum_{\mathcal{K}_k \subseteq \mathcal{K}_{K+1}} \int \xi(\mathbf{T}^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_{K+1}\setminus\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) dF(\mathbf{T}^{\mathcal{K}_k})$$

where the second equality follows from applying Lemma 1 to $\tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}})$ and the final equality from Lemma 2. This proves that the K-way AMIE is a linear function of the K-way AIEs.

We next prove that the K-way AIE can be written as a linear function of the K-way AMIEs. We will show this by mathematical induction. We already show the desired result holds for $K = 2$. Choose any $K \geq 2$ and assume that the following equality holds,

$$\xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K};\mathbf{t}_0^{\mathcal{K}_K}) \;=\; \sum_{k=1}^{K}(-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k})$$

Using the definition of the K-way AIE given in Lemma 1, we have

$$\begin{aligned}
\xi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}}) \;&=\; \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K};\mathbf{t}_0^{\mathcal{K}_K} \mid T_i^{K+1} = t^{K+1}) - \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K};\mathbf{t}_0^{\mathcal{K}_K} \mid T_i^{K+1} = t_0^{K+1}) \\
&= \sum_{k=1}^{K}(-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t^{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t^{K+1}) \\
&\quad - \sum_{k=1}^{K}(-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t_0^{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t_0^{K+1}),
\end{aligned}$$

where the second equality follows from the assumption. Let us consider the following decomposition.

$$\begin{aligned}
&\sum_{k=1}^{K+1}(-1)^{K-k+1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_{K+1}} \pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_{K+1}\setminus\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_{K+1}\setminus\mathcal{K}_k}) \\
&= \sum_{k=1}^{K}(-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t^{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t_0^{K+1}) + (-1)^K \pi_{\mathcal{K}_{K+1}}(\mathbf{t}_0^{\mathcal{K}_K}, t^{K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_0^{K+1}) \\
&\quad + \sum_{k=1}^{K}(-1)^{K-k+1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t_0^{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t_0^{K+1}), \qquad (16)
\end{aligned}$$

where the first and second terms together represent the cases with $K+1 \in \mathcal{K}_k$, while the third term corresponds to the cases with $K+1 \in \mathcal{K}_{K+1} \setminus \mathcal{K}_k$. Note that these two cases are mutually exclusive and exhaustive. Finally, note the following equality,

$$\sum_{k=1}^{K}(-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t^{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}, t^{K+1})$$

$$= \sum_{k=1}^{K}(-1)^{K-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_K}\pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k},t^{K+1};\mathbf{t}_0^{\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k},t_0^{K+1})+(-1)^K\pi_{\mathcal{K}_{K+1}}(\mathbf{t}_0^{\mathcal{K}_K},t^{K+1};\mathbf{t}_0^{\mathcal{K}_K},t_0^{K+1}) \tag{17}$$

Then, together with equations (16) and (17), we obtain,

$$\xi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}}) = \sum_{k=1}^{K+1}(-1)^{K-k+1}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_{K+1}}\pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_{K+1}\setminus\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_{K+1}\setminus\mathcal{K}_k})$$

Thus, the desired linear relationship holds for any $K \geq 2$. □

A.3 Proof of Theorem 2

To prove the invariance of the K-way AMIE, note that Lemma 4 implies,

$$\pi_{\mathcal{K}_K}(\mathbf{t};\mathbf{t}_0) - \pi_{\mathcal{K}_K}(\tilde{\mathbf{t}};\mathbf{t}_0) = \sum_{k=1}^{K}(-1)^{K-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_K}\tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\tilde{\mathbf{t}}^{\mathcal{K}_k}) \tag{18}$$

$$\pi_{\mathcal{K}_K}(\mathbf{t};\tilde{\mathbf{t}}_0) - \pi_{\mathcal{K}_K}(\tilde{\mathbf{t}};\tilde{\mathbf{t}}_0) = \sum_{k=1}^{K}(-1)^{K-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_K}\tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\tilde{\mathbf{t}}^{\mathcal{K}_k}) \tag{19}$$

Thus, the K-way AMIE is interval invariant. To prove the lack of invariance of the K-way AIE, note that according to Lemma 3, we can rewrite equation (10) as follows.

$$\begin{aligned}
& \sum_{k=1}^{K}(-1)^{K-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_K}\left\{\tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_K\setminus\mathcal{K}_k}=\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k})-\tau_{\mathcal{K}_k}(\tilde{\mathbf{t}}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_K\setminus\mathcal{K}_k}=\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k})\right\}\\
= & \sum_{k=1}^{K}(-1)^{K-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_K}\left\{\tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_K\setminus\mathcal{K}_k}=\tilde{\mathbf{t}}_0^{\mathcal{K}_K\setminus\mathcal{K}_k})-\tau_{\mathcal{K}_k}(\tilde{\mathbf{t}}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_K\setminus\mathcal{K}_k}=\tilde{\mathbf{t}}_0^{\mathcal{K}_K\setminus\mathcal{K}_k})\right\}
\end{aligned}$$

It is clear that this equality does not hold in general because the K-way conditional ACEs are conditioned on different treatment values. Thus, the K-way AIE is not interval invariant. □

A.4 Proof of Theorem 3

We use L to denote the objective function in equation (10). Since it is a convex optimization problem, it has one unique solution and the solution should satisfy the following equalities.

$$\frac{\partial L}{\partial \mu} = 0, \qquad \frac{\partial L}{\partial \beta_\ell^j} = 0 \quad \text{for all } j, \text{ and } \ell \in \{0, 1, \cdots, L_j - 1\},$$

$$\begin{aligned}
\frac{\partial L}{\partial \beta_{\ell,m}^{jj'}} &= 0, \quad \text{for all } j \neq j',\ \ell \in \{0,1,\cdots,L_j-1\} \text{ and } m \in \{0,1,\ldots,L_{j'}-1\}, \\
\frac{\partial L}{\partial \beta_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k}} &= 0 \quad \text{for all } \mathbf{t}^{\mathcal{K}_k},\ \text{and } \mathcal{K}_k \subset \mathcal{K}_J \text{ such that } k \geq 3 \qquad (20)
\end{aligned}$$

For the sake of simplicity, we introduce the following notation.

$$\mathcal{S}(\mathbf{t}^{\mathcal{K}_k}) \ \equiv\ \{i; \mathbf{T}_i^{\mathcal{K}_k} = \mathbf{t}^{\mathcal{K}_k}\},\quad N_{\mathbf{t}^{\mathcal{K}_k}} \ \equiv\ \sum_{i=1}^{n} \mathbf{1}\{\mathbf{T}_i^{\mathcal{K}_k} = \mathbf{t}^{\mathcal{K}_k}\},\quad \hat{\mathbb{E}}[Y_i \mid \mathbf{T}_i^{\mathcal{K}_k} = \mathbf{t}^{\mathcal{K}_k}] \ \equiv\ \frac{1}{N_{\mathbf{t}^{\mathcal{K}_k}}} \sum_{i \in \mathcal{S}(\mathbf{t}^{\mathcal{K}_k})} Y_i$$

Then, from $\frac{\partial L}{\partial \boldsymbol{\beta}_{\mathbf{t}^{\mathcal{K}_J}}^{\mathcal{K}_J}} = 0$ for all $\mathbf{t}^{\mathcal{K}_J}$,

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\beta}_{\mathbf{t}^{\mathcal{K}_J}}^{\mathcal{K}_J}} = \sum_{i \in \mathcal{S}(\mathbf{t}^{\mathcal{K}_k})} -2\Bigg(Y_i - \mu - \sum_{j=1}^{J}\sum_{\ell=0}^{L_j-1} \beta_\ell^j \mathbf{1}\{T_{ij}=\ell\} - \sum_{j=1}^{J-1}\sum_{j'>j}\sum_{\ell=0}^{L_j-1}\sum_{m=0}^{L_{j'}-1} \beta_{\ell m}^{jj'} \mathbf{1}\{T_{ij}=\ell, T_{ij'}=m\} \\
- \sum_{k=3}^{J} \sum_{\mathcal{K}_k \subset \mathcal{K}_J} \sum_{\mathbf{t}^{\mathcal{K}_k}} \beta_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k} \mathbf{1}\{\mathbf{T}_i^{\mathcal{K}_k} = \mathbf{t}^{\mathcal{K}_k}\}\Bigg) = 0 \qquad (21)
\end{aligned}$$

Therefore, for all $\mathbf{t}^{\mathcal{K}_J}$,

$$\hat{\mu} + \sum_{k=1}^{J} \sum_{\mathcal{K}_k \subset \mathcal{K}_J} \sum_{\mathbf{t}^{\mathcal{K}_k}} \beta_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k} \mathbf{1}\{\mathbf{t}^{\mathcal{K}_k} \subset \mathbf{t}^{\mathcal{K}_J}\} \ = \ \widehat{\mathbb{E}}[Y_i \mid \mathbf{T}_i^{\mathcal{K}_J} = \mathbf{t}^{\mathcal{K}_J}].$$

For the first-order effect, we can use the weighted zero-sum constraints for all factors except for the j th factor. In particular, for all j and $t_{j\ell} \in \mathbf{t}^{\mathcal{K}_J}$,

$$\begin{aligned}
&\sum_{j' \neq j} \sum_{\ell=0}^{L_{j'}-1} \prod_{t_{j'\ell} \in \mathbf{t}^{\mathcal{K}_J \setminus j}} \Pr(T_{ij'}=\ell) \Bigg\{\hat{\mu} + \sum_{k=1}^{J} \sum_{\mathcal{K}_k \subset \mathcal{K}_J} \sum_{\mathbf{t}^{\mathcal{K}_k}} \beta_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k} \mathbf{1}\{\mathbf{t}^{\mathcal{K}_k} \in \mathbf{t}^{\mathcal{K}_J}\}\Bigg\} \\
&= \sum_{j' \neq j} \sum_{\ell=0}^{L_{j'}-1} \prod_{t_{j'\ell} \in \mathbf{t}^{\mathcal{K}_J \setminus j}} \Pr(T_{ij'}=\ell) \ \ \widehat{\mathbb{E}}[Y_i \mid T_{ij}=\ell, \mathbf{T}_i^{\mathcal{K}_J \setminus j} = \mathbf{t}^{\mathcal{K}_J \setminus j}] \\
\Longleftrightarrow \quad \hat{\beta}_\ell^j \ &= \ \sum_{j' \neq j} \sum_{\ell=0}^{L_{j'}-1} \prod_{t_{j'\ell} \in \mathbf{t}^{\mathcal{K}_J \setminus j}} \Pr(T_{ij'}=\ell) \ \ \widehat{\mathbb{E}}[Y_i \mid T_{ij}=\ell, \mathbf{T}_i^{\mathcal{K}_J \setminus j} = \mathbf{t}^{\mathcal{K}_J \setminus j}] - \hat{\mu}
\end{aligned}$$

In general, for all $\mathbf{t}^{\mathcal{K}_k}, \mathcal{K}_k \subset \mathcal{K}_J$ and $k \geq 2$,

$$\begin{aligned}
\hat{\beta}_{\mathbf{t}^{\mathcal{K}_k}}^{\mathcal{K}_k} = \sum_{j' \in \mathcal{K}_J \setminus \mathcal{K}_k} \sum_{\ell=0}^{L_{j'}-1} \prod_{t_{j'\ell} \in \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k}} \Pr(T_{ij'}=\ell) \ \ \widehat{\mathbb{E}}[Y_i \mid \mathbf{T}_i^{\mathcal{K}_k} = \mathbf{t}^{\mathcal{K}_k}, \mathbf{T}_i^{\mathcal{K}_J \setminus \mathcal{K}_k} = \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k}] \\
- \sum_{\mathcal{K}_p \subset \mathcal{K}_k} \sum_{\mathbf{t}^{\mathcal{K}_p}} \mathbf{1}\{\mathbf{t}^{\mathcal{K}_p} \subset \mathbf{t}^{\mathcal{K}_k}\} \hat{\beta}_{\mathbf{t}^{\mathcal{K}_p}}^{\mathcal{K}_p} - \hat{\mu} \qquad (22)
\end{aligned}$$

In addition, $\hat{\mu}$ is given as follows.

$$\hat{\mu} \ = \ \sum_{j=1}^{K} \sum_{\ell=0}^{L_j-1} \prod_{t_{j\ell} \in \mathbf{t}^{\mathcal{K}_J}} \Pr(T_{ij}=\ell) \ \ \widehat{\mathbb{E}}[Y_i \mid \mathbf{T}_i^{\mathcal{K}_J} = \mathbf{t}^{\mathcal{K}_J}]$$

Therefore, $(\hat{\mu}, \hat{\boldsymbol{\beta}})$ is uniquely determined. To confirm this solution is the minimizer of the optimization problem, we check all the equality conditions. For all $\mathbf{t}^{\mathcal{K}_k}, \mathcal{K}_k \subset \mathcal{K}_J, j \in \mathcal{K}_k$ and $k \geq 1$,

$$
\begin{aligned}
& \sum_{\ell=0}^{L_j-1} \Pr(T_{ij} = \ell) \mathbf{1}\{t_j = \ell\} \hat{\beta}^{\mathcal{K}_k}_{\mathbf{t}^{\mathcal{K}_k}} \\
= \ & \sum_{\ell=0}^{L_j-1} \Pr(T_{ij} = \ell) \mathbf{1}\{t_j = \ell\} \sum_{j' \in \mathcal{K}_J \setminus \mathcal{K}_k} \sum_{\ell=0}^{L_{j'}-1} \prod_{t_{j'\ell} \in \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k}} \Pr(T_{ij'} = \ell) \ \widehat{\mathbb{E}}[Y_i \mid \mathbf{T}_i^{\mathcal{K}_k} = \mathbf{t}^{\mathcal{K}_k}, \mathbf{T}_i^{\mathcal{K}_J \setminus \mathcal{K}_k} = \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k}] \\
& \qquad - \sum_{\ell=0}^{L_j-1} \Pr(T_{ij} = \ell) \mathbf{1}\{t_j = \ell\} \sum_{\mathcal{K}_p \subset \mathcal{K}_k} \sum_{\mathbf{t}^{\mathcal{K}_p}} \mathbf{1}\{\mathbf{t}^{\mathcal{K}_p} \subset \mathbf{t}^{\mathcal{K}_k}\} \hat{\beta}^{\mathcal{K}_p}_{\mathbf{t}^{\mathcal{K}_p}} - \hat{\mu} \\
= \ & \sum_{j' \in \{j, \mathcal{K}_J \setminus \mathcal{K}_k\}} \sum_{\ell=0}^{L_{j'}-1} \prod_{t_{j'\ell} \in \mathbf{t}^{\{j, \mathcal{K}_J \setminus \mathcal{K}_k\}}} \Pr(T_{ij'} = \ell) \ \widehat{\mathbb{E}}[Y_i \mid \mathbf{T}_i^{\mathcal{K}_k \setminus j} = \mathbf{t}^{\mathcal{K}_k \setminus j}, \mathbf{T}_i^{\{j, \mathcal{K}_J \setminus \mathcal{K}_k\}} = \mathbf{t}^{\{j, \mathcal{K}_J \setminus \mathcal{K}_k\}}] \\
& \qquad - \sum_{\mathcal{K}_p \subseteq \mathcal{K}_k \setminus j} \sum_{\mathbf{t}^{\mathcal{K}_p}} \mathbf{1}\{\mathbf{t}^{\mathcal{K}_p} \subseteq \mathbf{t}^{\mathcal{K}_k \setminus j}\} \hat{\beta}^{\mathcal{K}_p}_{\mathbf{t}^{\mathcal{K}_p}} - \hat{\mu} \\
= \ & 0
\end{aligned}
$$

where the final equality comes from equation (22) for $\hat{\beta}^{\mathcal{K}_k \setminus j}_{\mathbf{t}^{\mathcal{K}_k \setminus j}}$.

Furthermore, equation (21) implies all other equalities in equation (20). Therefore, the solution (equation (22) and equation (23)) satisfies all the equality conditions. Finally, we show that these estimators are unbiased for the AMEs and the AMIEs. Since $\widehat{\mathbb{E}}[Y_i \mid \mathbf{T}_i^{\mathcal{K}_J} = \mathbf{t}^{\mathcal{K}_J}]$ is an unbiased estimator of $\mathbb{E}[Y_i(\mathbf{t}^{\mathcal{K}_J})]$,

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}^{\mathcal{K}_k}_{\mathbf{t}^{\mathcal{K}_k}}] &= \sum_{j' \in \mathcal{K}_J \setminus \mathcal{K}_k} \sum_{\ell=0}^{L_{j'}-1} \prod_{t_{j'\ell} \in \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k}} \Pr(T_{ij'} = \ell) \ \mathbb{E}[Y_i(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k})] \\
& \qquad - \sum_{\mathcal{K}_p \subset \mathcal{K}_k} \sum_{\mathbf{t}^{\mathcal{K}_p}} \mathbf{1}\{\mathbf{t}^{\mathcal{K}_p} \subset \mathbf{t}^{\mathcal{K}_k}\} \mathbb{E}[\hat{\beta}^{\mathcal{K}_p}_{\mathbf{t}^{\mathcal{K}_p}}] - \hat{\mu} \\
\mathbb{E}[\hat{\beta}^{\mathcal{K}_k}_{\mathbf{t}^{\mathcal{K}_k}} - \hat{\beta}^{\mathcal{K}_k}_{\mathbf{t}_0^{\mathcal{K}_k}}] &= \sum_{j' \in \mathcal{K}_J \setminus \mathcal{K}_k} \sum_{\ell=0}^{L_{j'}-1} \prod_{t_{j'\ell} \in \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k}} \Pr(T_{ij'} = \ell) \ \mathbb{E}[Y_i(\mathbf{t}^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k}) - Y_i(\mathbf{t}_0^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_J \setminus \mathcal{K}_k})] \\
& \qquad - \sum_{\mathcal{K}_p \subset \mathcal{K}_k} \sum_{\mathbf{t}^{\mathcal{K}_p}} \mathbf{1}\{\mathbf{t}^{\mathcal{K}_p} \subset \mathbf{t}^{\mathcal{K}_k}\} \mathbb{E}[\hat{\beta}^{\mathcal{K}_p}_{\mathbf{t}^{\mathcal{K}_p}} - \hat{\beta}^{\mathcal{K}_p}_{\mathbf{t}_0^{\mathcal{K}_p}}] \\
&= \pi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k})
\end{aligned}
$$

$\square$

B Supplementary Appendix: Proofs of Lemmas

For the sake of completeness, we prove all the lemmas used in the mathematical appendix above.

B.1 Proof of Lemma 1

To simplify the proof, we start from Lemma 1 and prove it is equivalent to Definition 3. We prove it by induction. Equation (3.2) shows this correspondence holds for $K = 2$. Next, choose any $K \geq 2$ and assume that this relationship holds. That is, we assume the following equality,

$$\xi_{\mathcal{K}_K}(\mathbf{t}; \mathbf{t}_0) \;=\; \tau_{\mathcal{K}_K}(\mathbf{t}; \mathbf{t}_0) \;-\; \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \tag{23}$$

where the second summation is taken over all possible $\mathcal{K}_k \subseteq \mathcal{K}_K = \{1, \ldots, K\}$ with $|\mathcal{K}_k| = k$.

Using the definition of the K-way AIE in Lemma 1, we have,

$$\begin{aligned}
&\sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\{\mathcal{K}_k, K+1\}}(\mathbf{t}^{\mathcal{K}_k}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \\
= &\sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid T_{i,K+1} = t_{K+1}, \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \\
&\quad - \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid T_{i,K+1} = t_{0,K+1}, \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}),
\end{aligned} \tag{24}$$

where $\xi_{\{\mathcal{K}_k, K+1\}}(\mathbf{t}^{\mathcal{K}_k}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k})$ denote the conditional $(k+1)$-way AIE that includes the set of k treatments, $\mathcal{K}_k$, as well as the $(K+1)$th treatment while fixing $\overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k}$ to $\mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}$. Therefore, we have,

$$\begin{aligned}
&\xi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) \\
= \;& \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K} \mid T_{i,K+1} = t_{K+1}) - \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K} \mid T_{i,K+1} = t_{0,K+1}) \\
= \;& \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_{K+1}) - \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K}, t_{0,K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_{0,K+1})
\end{aligned}$$

$$
\begin{aligned}
&\quad - \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\{\mathcal{K}_k, K+1\}}(\mathbf{t}^{\mathcal{K}_k}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \\
&= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_{K+1}) - \sum_{k=1}^{K} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}) \\
&\quad - \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\{\mathcal{K}_k, K+1\}}(\mathbf{t}^{\mathcal{K}_k}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \qquad (25)
\end{aligned}
$$

where the second equality follows from equation (24), and the third equality is based on the application of the assumption given in equation (23) while conditioning on $T_{i,K+1} = t_{0,K+1}$.

Next, consider the following decomposition,

$$
\begin{aligned}
&\sum_{k=1}^{K} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_{K+1}} \xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}) \\
&= \sum_{k=1}^{K} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}) \\
&\quad + \sum_{k=1}^{K-1} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \xi_{\{\mathcal{K}_k, K+1\}}(\mathbf{t}^{\mathcal{K}_k}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \\
&\quad + \xi_{(K+1)}(t_{K+1}; t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K} = \mathbf{t}_0^{\mathcal{K}_K}) \qquad (26)
\end{aligned}
$$

where the first term corresponds to the cases with $K+1 \in \mathcal{K}_{K+1} \setminus \mathcal{K}_k$, while the second and third terms together represent the cases with $K+1 \in \mathcal{K}_k$. Note that these two cases are mutually exclusive and exhaustive. Finally, note the following equality,

$$
\tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_{K+1}) = \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) - \xi_{(K+1)}(t_{K+1}; t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K} = \mathbf{t}_0^{\mathcal{K}_K}).
$$

Then, together with equations (25) and (26), we obtain, the desired result,

$$
\xi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) = \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) - \sum_{k=1}^{K} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_{K+1}} \xi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k})
$$

Thus, the lemma holds for any $K \geq 2$. $\square$

B.2 Proof of Lemma 2

To begin, we prove the following equality by mathematical induction.

$$
\begin{aligned}
&\xi_{\mathcal{K}_K}(\tilde{\mathbf{T}}^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_K}) \\
&= \xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k} \mid \tilde{\mathbf{T}}^{\mathcal{K}_k}) + \sum_{\ell=1}^{k}(-1)^\ell \sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k} \xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k} \mid \tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell})
\end{aligned}
\tag{27}
$$

First, it is clear that this equality holds when $k = 1$. That is, for a given $\mathcal{K}_1$, we have,

$$
\begin{aligned}
&\xi_{\mathcal{K}_K}(\tilde{T}^{\mathcal{K}_1}, \mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_1}; \mathbf{t}_0^{\mathcal{K}_K}) \\
&= \xi_{\mathcal{K}_K\setminus\mathcal{K}_1}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_1}; \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_1} \mid \tilde{T}_i^{\mathcal{K}_1} = \tilde{t}^{\mathcal{K}_1}) - \xi_{\mathcal{K}_K\setminus\mathcal{K}_1}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_1}; \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_1} \mid T_i^{\mathcal{K}_1} = t_0^{\mathcal{K}_1})
\end{aligned}
\tag{28}
$$

Now, assume that the equality holds for k. Without loss of generality, we suppose $\mathcal{K}_k = \{1, 2, \cdots, k\}$ and $\mathcal{K}_{k+1} = \{1, 2, \cdots, k, k+1\}$. By the definition of the K-way AIE,

$$
\begin{aligned}
&\xi_{\mathcal{K}_K}(\tilde{\mathbf{T}}^{\mathcal{K}_{k+1}}, \mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}; \mathbf{t}_0^{\mathcal{K}_K}) \\
&= \xi_{\mathcal{K}_K\setminus(k+1)}(\tilde{\mathbf{T}}^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}; \mathbf{t}_0^{\mathcal{K}_K\setminus(k+1)} \mid \tilde{T}_i^{k+1}) - \xi_{\mathcal{K}_K\setminus(k+1)}(\tilde{\mathbf{T}}^{\mathcal{K}_k}, \mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}; \mathbf{t}_0^{\mathcal{K}_K\setminus(k+1)} \mid T_i^{k+1} = t_0^{k+1}) \\
&= \xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}; \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}} \mid \tilde{\mathbf{T}}_i^{\mathcal{K}_{k+1}}) \\
&\quad + \sum_{\ell=1}^{k}(-1)^\ell \sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k} \xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}} \mid \tilde{\mathbf{T}}^{\mathcal{K}_{k+1}\setminus\mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell}) \\
&\quad - \xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}; \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}} \mid \tilde{\mathbf{T}}_i^{\mathcal{K}_k}, T_i^{k+1} = t_0^{k+1}) \\
&\quad + \sum_{\ell=1}^{k}(-1)^{\ell+1} \sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k} \xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}} \mid \tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell}, T_i^{k+1} = t_0^{k+1}),
\end{aligned}
\tag{29}
$$

where the second equality follows from the assumption.

Next, consider the following decomposition.

$$
\sum_{\ell=1}^{k+1}(-1)^\ell \sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_{k+1}} \xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}, \mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}} \mid \tilde{\mathbf{T}}^{\mathcal{K}_{k+1}\setminus\mathcal{K}_\ell}, \overline{\mathbf{T}}_i^{\mathcal{K}_\ell} = \mathbf{t}_0^{\mathcal{K}_\ell})
$$

$$
\begin{aligned}
&= \sum_{\ell=1}^{k}(-1)^{\ell}\sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k}\xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}\mid\tilde{\mathbf{T}}^{\mathcal{K}_{k+1}\setminus\mathcal{K}_\ell},\overline{\mathbf{T}}_i^{\mathcal{K}_\ell}=\mathbf{t}_0^{\mathcal{K}_\ell})\\
&\quad-\xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}};\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}\mid\tilde{\mathbf{T}}_i^{\mathcal{K}_k},T_i^{k+1}=t_0^{k+1})\\
&\quad+\sum_{\ell=1}^{k}(-1)^{\ell+1}\sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k}\xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}\mid\tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell},\overline{\mathbf{T}}_i^{\mathcal{K}_\ell}=\mathbf{t}_0^{\mathcal{K}_\ell},T_i^{k+1}=t_0^{k+1}),
\end{aligned}
\tag{30}
$$

where the first term corresponds to the case in which $\mathcal{K}_\ell\subseteq\mathcal{K}_{k+1}$ in the left side of the equation does not include the $(k+1)$th treatment, and the second and third terms jointly express the case in which $\mathcal{K}_\ell\subseteq\mathcal{K}_{k+1}$ in the left side of the equation does include the $(k+1)$th treatment.

Putting together equations (29) and (30), we have,

$$
\begin{aligned}
&\xi_{\mathcal{K}_K}(\tilde{\mathbf{T}}^{\mathcal{K}_{k+1}},\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_K})\\
&=\xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}\mid\tilde{\mathbf{T}}^{\mathcal{K}_{k+1}})\\
&\quad+\sum_{\ell=1}^{k+1}(-1)^{\ell}\sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_{k+1}}\xi_{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_{k+1}}\mid\tilde{\mathbf{T}}^{\mathcal{K}_{k+1}\setminus\mathcal{K}_\ell},\overline{\mathbf{T}}_i^{\mathcal{K}_\ell}=\mathbf{t}_0^{\mathcal{K}_\ell}).
\end{aligned}
$$

Therefore, equation (27) holds in general. Finally, under Assumption 2,

$$
\begin{aligned}
&\int_{\overline{\mathcal{F}}^{\mathcal{K}_k}}\xi_{\mathcal{K}_K}(\tilde{\mathbf{T}}^{\mathcal{K}_k},\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_K})dF(\tilde{\mathbf{T}}^{\mathcal{K}_k})\\
&=\int_{\overline{\mathcal{F}}^{\mathcal{K}_k}}\xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}\mid\tilde{\mathbf{T}}^{\mathcal{K}_k})dF(\tilde{\mathbf{T}}^{\mathcal{K}_k})\\
&\quad+\sum_{\ell=1}^{k}(-1)^{\ell}\sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k}\int_{\overline{\mathcal{F}}^{\mathcal{K}_k}}\xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}\mid\tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell},\overline{\mathbf{T}}_i^{\mathcal{K}_\ell}=\mathbf{t}_0^{\mathcal{K}_\ell})dF(\tilde{\mathbf{T}}^{\mathcal{K}_k})\\
&=\xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k})\\
&+\sum_{\ell=1}^{k}(-1)^{\ell}\sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k}\left\{\int_{\overline{\mathcal{F}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell}}\int_{\overline{\mathcal{F}}^{\mathcal{K}_\ell}}\xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}\mid\tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell},\overline{\mathbf{T}}_i^{\mathcal{K}_\ell}=\mathbf{t}_0^{\mathcal{K}_\ell})dF(\tilde{\mathbf{T}}^{\mathcal{K}_\ell}\mid\tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell})dF(\tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell})\right\}\\
&=\xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k})\\
&\quad+\sum_{\ell=1}^{k}(-1)^{\ell}\sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k}\int_{\overline{\mathcal{F}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell}}\xi_{\mathcal{K}_K\setminus\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_K\setminus\mathcal{K}_k},\mathbf{t}_0^{\mathcal{K}_K\setminus\mathcal{K}_k}\mid\tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell},\overline{\mathbf{T}}_i^{\mathcal{K}_\ell}=\mathbf{t}_0^{\mathcal{K}_\ell})dF(\tilde{\mathbf{T}}^{\mathcal{K}_k\setminus\mathcal{K}_\ell})
\end{aligned}
$$

This completes the proof of Lemma 2. □

B.3 Proof of Lemma 3

We prove the lemma by induction. For $K = 2$, equation (3.2) shows that the lemma holds. Choose any $K \geq 2$ and assume that the lemma holds for all k with $1 \leq k \leq K$. Then,

$$
\begin{aligned}
& \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K} \mid T_{i,K+1} = t_{K+1}) \\
&= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_{K+1}) + \sum_{k=1}^{K-1} (-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, t_{K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \\
&= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_{K+1}) \\
&\quad + \sum_{k=1}^{K-1} (-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \Bigg[\tau_{\{\mathcal{K}_k, K+1\}}(\mathbf{t}^{\mathcal{K}_k}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \\
&\qquad\qquad - \tau_{K+1}(t_{K+1}; t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_{K+1} \setminus \ (K+1)} = \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \ (K+1)}) \Bigg] \\
&= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_{K+1}) \\
&\quad + \sum_{k=1}^{K-1} (-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \tau_{\{\mathcal{K}_k, K+1\}}(\mathbf{t}^{\mathcal{K}_k}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_k}, t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_K \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_K \setminus \mathcal{K}_k}) \\
&\quad + \sum_{k=1}^{K-1} (-1)^{K-k+1} \binom{K}{k} \tau_{K+1}(t_{K+1}; t_{0,K+1} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_{K+1} \setminus \ (K+1)} = \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \ (K+1)}).
\end{aligned}
\tag{31}
$$

Next, note the following decomposition,

$$
\begin{aligned}
\xi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) &= \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K} \mid T_{i,K+1} = t_{K+1}) - \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_L} \mid T_{i,K+1} = t_{0,K+1}) \\
&= \xi_{\mathcal{K}_K}(\mathbf{t}^{\mathcal{K}_K}; \mathbf{t}_0^{\mathcal{K}_K} \mid T_{i,K+1} = t_{K+1}) \\
&\quad - \sum_{k=1}^{K} (-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}),
\end{aligned}
$$

Substituting equation (31) into this equation, we obtain

$$
\begin{aligned}
& \xi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}}; \mathbf{t}_0^{\mathcal{K}_{K+1}}) \\
&= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K}, t_{K+1}; \mathbf{t}_0^{\mathcal{K}_K}, t_{K+1}) - \sum_{k=1}^{K} (-1)^{K-k} \sum_{\mathcal{K}_k \subseteq \mathcal{K}_K} \tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k}; \mathbf{t}_0^{\mathcal{K}_k} \mid \overline{\mathbf{T}}_i^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k} = \mathbf{t}_0^{\mathcal{K}_{K+1} \setminus \mathcal{K}_k}),
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{k=1}^{K-1}(-1)^{K-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_K}\tau_{\{\mathcal{K}_k,K+1\}}(\mathbf{t}^{\mathcal{K}_k},t_{K+1};\mathbf{t}_0^{\mathcal{K}_k},t_{0,K+1}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_K\backslash\mathcal{K}_k}=\mathbf{t}_0^{\mathcal{K}_K\backslash\mathcal{K}_k})\\
&+\sum_{k=1}^{K-1}(-1)^{K-k+1}\binom{K}{k}\tau_{K+1}(t_{K+1};t_{0,K+1}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\ (K+1)}=\mathbf{t}_0^{\mathcal{K}_{K+1}\backslash\ (K+1)})\\
=\ &\tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_K},t_{K+1};\mathbf{t}_0^{\mathcal{K}_K},t_{K+1})-\sum_{k=2}^{K}(-1)^{K-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_{K+1}}\tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\mathcal{K}_k}=\mathbf{t}_0^{\mathcal{K}_{K+1}\backslash\mathcal{K}_k}),\\
&+(-1)^K\sum_{\mathcal{K}_1\subseteq\mathcal{K}_K}\tau_{\mathcal{K}_1}(t^{\mathcal{K}_1},t_0^{\mathcal{K}_1}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\ \mathcal{K}_1}=\mathbf{t}_0^{\mathcal{K}_{L+1}\backslash\ \mathcal{K}_1})\\
&+\sum_{k=1}^{K-1}(-1)^{K-k+1}\binom{K}{k}\tau_{K+1}(t_{K+1};t_{0,K+1}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\ (K+1)}=\mathbf{t}_0^{\mathcal{K}_{K+1}\backslash\ (K+1)})\\
=\ &\tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}})-\tau_{K+1}(t_{K+1};t_{0,K+1}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\ (K+1)}=\mathbf{t}_0^{\mathcal{K}_{K+1}\backslash\ (K+1)})\\
&-\sum_{k=2}^{K}(-1)^{K-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_{K+1}}\tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\mathcal{K}_k}=\mathbf{t}_0^{\mathcal{K}_{K+k}\backslash\mathcal{K}_k}),\\
&+(-1)^K\sum_{\mathcal{K}_1\subseteq\mathcal{K}_K}\tau_{\mathcal{K}_1}(t^{\mathcal{K}_1},t_0^{\mathcal{K}_1}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\ \mathcal{K}_1}=\mathbf{t}_0^{\mathcal{K}_{K+1}\backslash\ \mathcal{K}_1})\\
&+\sum_{k=1}^{K-1}(-1)^{K-k+1}\binom{K}{k}\tau_{K+1}(t_{K+1};t_{0,K+1}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\ (K+1)}=\mathbf{t}_0^{\mathcal{K}_{K+1}\backslash\ (K+1)})\\
=\ &\sum_{k=1}^{K+1}(-1)^{K-k+1}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_{K+1}}\tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}\mid\overline{\mathbf{T}}_i^{\mathcal{K}_{K+1}\backslash\mathcal{K}_k}=\mathbf{t}_0^{\mathcal{K}_{K+1}\backslash\mathcal{K}_k}),
\end{aligned}
$$

where the final equality follows because

$$-1+\sum_{k=1}^{K-1}(-1)^{K-k+1}\binom{K}{k}\ =\ (-1)^K.$$

Thus, by induction, the theorem holds for any $K\geq 2$. □

B.4 Proof of Lemma 4

We prove the lemma by induction. For $K=2$, equation (3.2) shows this theorem holds. Choose any $K\geq 2$ and assume that the lemma holds for all k with $1\leq k\leq K$. That is, let $\mathcal{K}_k\subseteq\mathcal{K}_K=\{1,\ldots,K\}$ with $|\mathcal{K}_k|=k$ where $k=1,\ldots,K$, and assume the following equality,

$$\pi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k})\ =\ \sum_{\ell=1}^{k}(-1)^{k-\ell}\sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k}\tau_{\mathcal{K}_\ell}(\mathbf{t}^{\mathcal{K}_\ell};\mathbf{t}_0^{\mathcal{K}_\ell}).$$

Using this assumption as well as the definition of the K-way AMIE given in Definition 2, we have,

$$
\begin{aligned}
\pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}}) &= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}}) - \sum_{k=1}^{K}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_{K+1}} \pi_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}) \\
&= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}}) + \sum_{k=1}^{K}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_{K+1}} \sum_{\ell=1}^{k}(-1)^{k+1-\ell}\sum_{\mathcal{K}_\ell\subseteq\mathcal{K}_k} \tau_{\mathcal{K}_\ell}(\mathbf{t}^{\mathcal{K}_\ell};\mathbf{t}_0^{\mathcal{K}_\ell})
\end{aligned} \tag{32}
$$

Next, we determine the coefficient for $\tau_{\mathcal{K}_m}(\mathbf{t}^{\mathcal{K}_m};\mathbf{t}_0^{\mathcal{K}_m})$ in the second term of equation (32) for each m with $1 \leq m \leq K$. Note that $\tau_{\mathcal{K}_m}(\mathbf{t}^{\mathcal{K}_m};\mathbf{t}_0^{\mathcal{K}_m})$ would not appear in this term if $m > k$. That is, for a given m, we only need to consider the cases where the index for the first summation satisfies $m \leq k \leq K$. Furthermore, for any given such k, there exist $\binom{K+1-m}{k-m}$ ways to choose $\mathcal{K}_k$ in the second summation such that $\mathcal{K}_m \subseteq \mathcal{K}_k$. Once such $\mathcal{K}_k$ is selected, $\mathcal{K}_m$ appears only once in the third and fourth summations together and is multiplied by $(-1)^{k+1-m}$. Therefore, the coefficient for $\tau_{\mathcal{K}_m}(\mathbf{t}^{\mathcal{K}_m};\mathbf{t}_0^{\mathcal{K}_m})$ is equal to,

$$
\sum_{k=m}^{K}(-1)^{k+1-m}\binom{K+1-m}{k-m} \;=\; (-1)^{K+1-m}.
$$

Putting all of these together,

$$
\begin{aligned}
\pi_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}}) &= \tau_{\mathcal{K}_{K+1}}(\mathbf{t}^{\mathcal{K}_{K+1}};\mathbf{t}_0^{\mathcal{K}_{K+1}}) + \sum_{k=1}^{K}(-1)^{K+1-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_{K+1}} \tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k}) \\
&= \sum_{k=1}^{K+1}(-1)^{K+1-k}\sum_{\mathcal{K}_k\subseteq\mathcal{K}_{K+1}} \tau_{\mathcal{K}_k}(\mathbf{t}^{\mathcal{K}_k};\mathbf{t}_0^{\mathcal{K}_k})
\end{aligned}
$$

Since the theorem holds for $K+1$, we have shown that it holds for any $K \geq 2$. $\square$