Using the **Reconcile Algorithm**, we address discrepancies between **Conditional Average Treatment Effect (CATE) Estimators** to solve the **reference class problem** in causal inference for more consistent individual predictions.

Reconciling Heterogeneous Effects in Causal Inference

Audrey Chang[1] (audreychang@college.harvard.edu), Alexander Tolbert[2], Emily Diana[3]

1. Harvard University, 2. Emory University, 3. Toyota Technological Institute at Chicago, Carnegie Mellon University

Background

- *Reference class problem:* choice of reference class leads to varied predictions, from limited data
- *Predictive multiplicity problem:* models achieve similar aggregate performance but differ in individual predictions
- *Reconcile Algorithm:* reconciles models to solve for predictive multiplicity, similar to multicalibration algorithm

Research objectives

- Explain equivalence between predictive multiplicity and reference class problem
- Solve reference class problem in causal inference by applying the Reconcile Algorithm to reconcile CATE Estimators

The Reference Class Problem across domains

Philosophy — ***Reference Class Problem***

- Source of stereotyping and bias (Cosmos Club anecdote)

Definition. *Reference Class Problem.* In a universe $\mathcal{X}$ with a distribution over elements e and a probabilistic predicate $P : \mathcal{X} \to \{0,1\}$, the individual reference class problem arises for e if there are distinct reference classes S_i and S_j such that (1) $e \in S_i$ and $e \in S_j$, (2) $S_i \neq S_j$, and (3) $S_i \not\subseteq S_j$, $S_j \not\subseteq S_i$. Furthermore, if $\Pr[P(e)|e \in S_i] \neq \Pr[P(e)|e \in S_j]$, this results in ambiguity in assigning a definitive probability to $P(e)$ based on a single reference class.

Statistics — ***Heterogeneous Effects***

Under heterogeneous treatment effects,

$$Y(1)_i - Y(0)_i \neq Y(1)_j - Y(0)_j$$

predict individualized outcome:

$$CATE(X = x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

On what basis to choose X?

Computer Science — ***Predictive Multiplicity***

- Reconcile Algorithm reconciles individual probability predictions

Definition. *Predictive Multiplicity.* Given a baseline classifier h_0 and an error tolerance ϵ, a prediction problem exhibits predictive multiplicity over the ϵ-level set $S_\epsilon(h_0)$ if there exists a classifier $h \in S_\epsilon(h_0)$ such that $h(x_i) \neq h_0(x_i)$ for some instance x_i in the training dataset, indicating that competing models within the set achieve similar aggregate performance while differing in their individual predictions.

Dawid's Insight

Two approaches to individual prediction:

Individual to Group (i2G)

- Model individual-level predictions before aggregating to groups
- Calculate group probabilities by averaging within a reference class and compare to empirical averages to tune/falsify predictions
- Non-unique models → *predictive multiplicity problem!*

Group to Individual (G2i)

- Start with aggregate data to derive individual predictions by selecting a sufficiently large reference class and using its proportion for estimates.
- How to select reference class? → *reference class problem*
  - Inputs may belong to multiple reference classes; can't condition on all of them
  - Different choices of reference class lead to different estimates

Equivalence!

- Both originate from data failing to encode unique estimates for individual probabilities
  - **G2i** - data samples make it difficult after conditioning, curse of dimensionality
  - **i2G** - challenging to generalize reference classes to make confident individual predictions
- Equivalent goal and limitation: data evidences multiple possible "true probabilities"
- No longer *choosing* ref class (choosing between CATE estimators) but *reconciling* CATE estimators via subgroup performance

Reconcile Algorithm

Roth, Tolbert, and Weinstein propose the **Reconcile Algorithm** to solve predictive multiplicity / reference class problem.

1. **Contest** a model f_A with another model f_B if they disagree substantially on individual predictions. Extract a large reference class S from their disagreement region. At least one of the models has a lower mean squared error, and thus falsifies the other model.
2. **Update** the falsified model (WLOG) f_A to produce a new model f'_A that makes predictions that are correct on average over S. f'_A is now not falsified and more accurate.
3. **Repeat** steps 1-2 until f_A and f_B agree within some error bound.

The algorithm can be applied to any model f.

→ We apply Reconcile to address CATE estimator disagreement, adopting the i2G perspective: begin modeling individual treatment effects, rather than starting from group-level averages, and falsify to group-level statistics!

A simple reduction

Construct a new variable, estimated treatment effect:

$$\hat{y} = \mathbb{E}[y \mid X = x, T = 1] - \mathbb{E}[y \mid X = x, T = 0].$$

This makes our outcome continuous instead of discrete (necessary for Reconcile). Then, define the loss function over which Reconcile will minimize. We use Brier loss.

$$B(\hat{\tau}, \mathcal{D}) = \mathbb{E}_{(x,y,t)\sim\mathcal{D}}\left[(\hat{\tau}(x) - (\mathbb{E}[y|X = x, T = 1] - \mathbb{E}[y|X = x, T = 0]))^2\right]$$

Hence we'll use the following group-level statistic to falsify our predictions:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}(x_i) - \hat{y}_i)^2,$$

Note that this is equivalent to the Expected Mean Squared Error, a commonly-used metric for assessing CATE performance. So applying Reconcile can yield a unified model that optimally estimates the CATE.

Expanded for clarity:

Algorithm 1: ReconcileCATE

Let $t = t_1 = t_2 = 0$ and $\hat{\tau}_1^{t_1} = \hat{\tau}_1, \hat{\tau}_2^{t_2} = \hat{\tau}_2$

Let $m = \left\lceil \frac{2}{\sqrt{\alpha\epsilon}} \right\rceil$

while $\mu(U_\epsilon(\hat{\tau}_1^{t_1}, \hat{\tau}_2^{t_2})) \geq \alpha$ **do**

  for *each* $\bullet \in \{>, <\}$ *and* $i \in \{1, 2\}$ **do**

    $v_*^\bullet = \mathbb{E}_{(x,y,t\sim D)}[y|x \in U_\epsilon^\bullet(\hat{\tau}_1^{t_1}, \hat{\tau}_2^{t_2}), T = 1] - \mathbb{E}_{(x,y,t\sim D)}[y|x \in U_\epsilon^\bullet(\hat{\tau}_1^{t_1}, \hat{\tau}_2^{t_2}), T = 0]$

    $v_i^\bullet = \mathbb{E}_{(x,y,t\sim D)}[\hat{\tau}_i(x)^{t_i}|x \in U_\epsilon^\bullet(\hat{\tau}_1^{t_1}, \hat{\tau}_2^{t_2})]$

  Let

$$(i_t, \bullet_t) = \underset{\text{arg max}}{i \in \{1,2\}, \bullet \in \{>,<\}} \mu(U_\epsilon(\hat{\tau}_1^{t_1}, \hat{\tau}_2^{t_2})) \cdot (v_*^\bullet - v_i^\bullet)^2$$

  and

$$g_t(x) = \begin{cases} 1 & x \in U^{\bullet_t}(\hat{\tau}_1^{t_1}, \hat{\tau}_2^{t_2}) \\ 0 & \text{otherwise} \end{cases}$$

  Let

$$\tilde{\Delta}_t = \left(\mathbb{E}_{(x,y,t)\sim\mathcal{D}}[y|g_t(x) = 1, T = 1] - \mathbb{E}_{(x,y,t)\sim\mathcal{D}}[y|g_t(x) = 1, T = 0]\right) - \mathbb{E}_{(x,y,t)\sim\mathcal{D}}[\hat{\tau}_{i_t}^{t_i}(x)|g_t(x) = 1]$$

$$\Delta_t = \text{Round}(\tilde{\Delta}_t, m)$$

  Let $\hat{\tau}_i^{t_i+1}(x) = h(x, \hat{\tau}_i^{t_i}, g_t, \Delta_t)$, $t_i = t_i + 1, t = t + 1$

  Output $(\hat{\tau}_1^{t_1}, \hat{\tau}_2^{t_2})$.

EAAMO HARVARD UNIVERSITY Carnegie Mellon University EMORY UNIVERSITY Diana-Tolbert AI & Philosophy Lab