Personalized Treatment Selection using Causal Heterogeneity

Kinjal Basu

Sr. Staff Researcher @ LinkedIn AI

INFORMS 2021

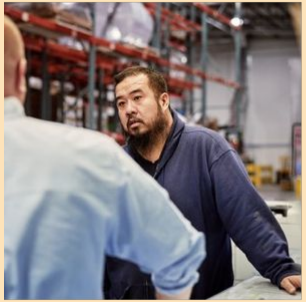
Joint work with Ye Tu, Preetam Nandy, Cyrus DiCiccio, Romil Bansal, Padmini Jaikumar & Shaunak Chatterjee

Tu et. al (2021)

Overview

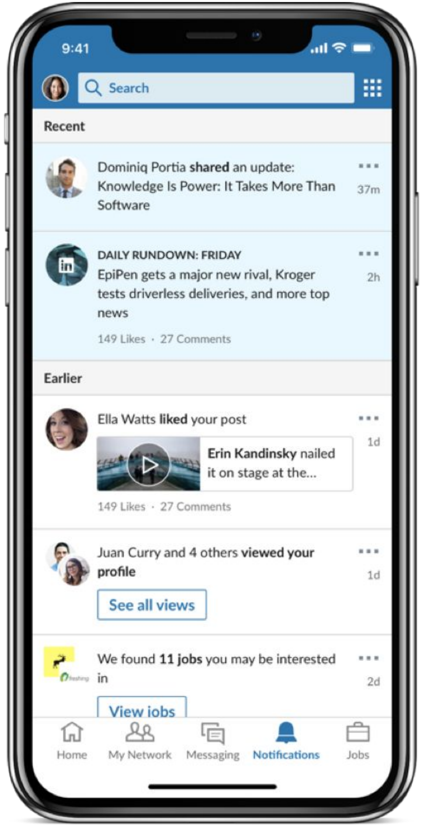
Introduction






Example Problem: Notification Systems



- Notifications are an important driver for **member visits** and **engagement**.
- Sending more notifications can increase visits, but it also has **negative consequences** where members can disable the system.
- **Challenging Business Problem**: How to identify the optimal number of relevant notifications that we should send to our member?

Heterogeneity of Treatment Effects

Randomized experimentation (**A/B testing**) is widely used in the internet industry to measure the metric impact obtained by different **treatment variants**.

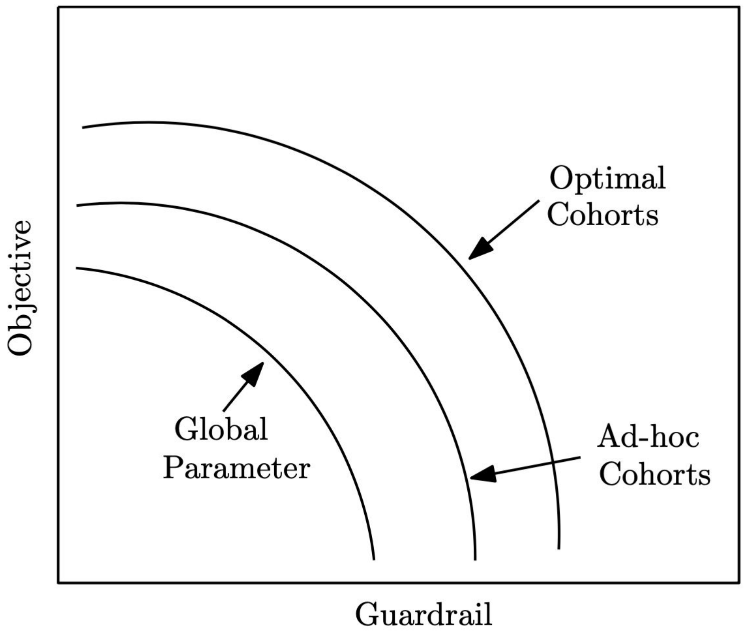
Treatment example: different number of notifications that members receive

The effect of a given treatment can be **heterogeneous** across experimental units.

Personalized Treatment Selection

Global allocation: identify the treatment variant that performs the best in the entire population and ramp that variant to everyone.

A **personalized approach** for treatment selection can greatly improve upon the usual global selection strategy.



Major Contributions

A **general framework** for identifying the optimal cohorts and allocating the optimal treatment per cohort.

- **Framework of solutions**: With guidance on which one to pick and when
- **Technical novelty**
  - Merging tree algorithm - Selects optimal cohorts
  - Multiple cooperative stochastic approximation - Selecting optimal treatments
- **Real-world application**
  - Building the serving infrastructure
  - Strong, positive results from a large scale industrial application

Problem Set-up

Notations and Objective

Let k = 0 denote the main success metric

- Sessions in the notifications example

Other constraints metrics k = 1,2..,K

- Clicks
- Disables, etc

Maximize Sessions
S.t. Clicks > c
Disables < d

Problem variable (x) :

- Numbers of notifications we sent to members

Notations and Objective

Objective 1: Find optimal cohorts

Symbol	Meaning
J	Total number of treatment variants or choices.
K	Total number of guardrail metrics
C_i	i-th cohort (the smallest cohort would be a in-dividual member) for $i = 1, \ldots, n$.

Notations and Objective

Objective 1: Find optimal cohorts

Symbol	Meaning
J	Total number of treatment variants or choices.
K	Total number of guardrail metrics
C_i	i-th cohort (the smallest cohort would be a individual member) for $i = 1, \ldots, n$.
$\mathbf{U}_k$	Vectorized version of $U_{i,j}^k$, which is the causal effect in metric k by variant j in cohort C_i.
$\boldsymbol{\mu}_k$	Mean of $\mathbf{U}_k$
$\boldsymbol{\sigma}_k^2\mathbf{I}$	Variance of $\mathbf{U}_k$

Notations and Objective

Objective 1: Find optimal cohorts

Objective 2: Allocate the optimal treatment to the cohorts

Symbol	Meaning
J	Total number of treatment variants or choices.
K	Total number of guardrail metrics
C_i	i-th cohort (the smallest cohort would be a individual member) for $i = 1, \ldots, n$.
$\mathbf{U}_k$	Vectorized version of $U^k_{i,j}$, which is the causal effect in metric k by variant j in cohort C_i.
$\boldsymbol{\mu}_k$	Mean of $\mathbf{U}_k$
$\boldsymbol{\sigma}_k^2\mathbf{I}$	Variance of $\mathbf{U}_k$
$\mathbf{x}$	The assignment vector.

Notations and Objective

Objective 1: Find optimal cohorts

Objective 2: Allocate the optimal treatment to the cohorts

Formally, we wish to get the optimal x* by solving the following:

$$\begin{aligned} \underset{\mathbf{x}}{\text{Maximize}} \quad & \mathbf{x}^T \mathbf{U}_0 \\ \text{subject to} \quad & \mathbf{x}^T \mathbf{U}_k \leq c_k \qquad \text{for } k = 1, \ldots, K. \\ & \sum_j x_{i,j} = 1 \ \ \forall i, \qquad 0 \leq \mathbf{x} \leq 1 \end{aligned}$$

Symbol	Meaning
J	Total number of treatment variants or choices.
K	Total number of guardrail metrics
C_i	i-th cohort (the smallest cohort would be a individual member) for $i = 1, \ldots, n$.
$\mathbf{U}_k$	Vectorized version of $U^k_{i,j}$, which is the causal effect in metric k by variant j in cohort C_i.
$\boldsymbol{\mu}_k$	Mean of $\mathbf{U}_k$
$\boldsymbol{\sigma}_k^2\mathbf{I}$	Variance of $\mathbf{U}_k$
$\mathbf{x}$	The assignment vector.

Problem Breakdown

1. Identify member cohorts $C_1, \ldots, C_n$ using data from randomized experiments to estimate the causal effect U_k for each cohort

2. Optimally allocate treatment variants x* to each member cohort by solving the optimization problem.

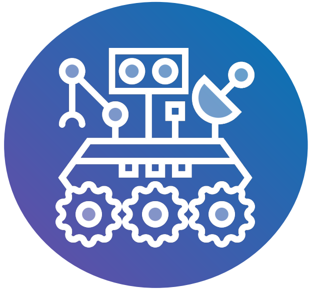
Methodology

Cohort-Level Heterogeneity

We use the recursive partitioning technique from Athey and Imbens [1] to identify the heterogeneous cohorts.

Regression Tree (CART)

- Predict Y
- Splitting Objective: MSE(Y)



Causal Tree

- Estimate treatment effect τ E(Y(1))-E(Y(0))
- Splitting Objective: MSE(τ) + Variance regularizer

Multiple treatments and metrics

Causal tree can only handle one objective metric and a binary treatment definition at a time.

- One option could be merging the $J\ (K + 1)$ tree models into one single cohort assignment.
- Simply merging all the trees would fragment the cohorts into very small subsets with **extremely noisy estimations**.
- We avoid this unwanted noise by carefully exploiting the within cohort homogeneity of the treatment effect by **Algorithm 1**.

Merging Trees - Algorithm 1

We sequentially merge the cohort sets $S_{j,k} = \{ C_1^{j,k}, \ldots, C_n^{j,k} \}$

For each treatment j and each metric k, we **retain** the estimated treatment effect and its variance from the **original cohort**.

Since each $S_{j,k}$ is exhaustive, this provides estimates of treatment effect and its variance for all sub-partitions.

Algorithm 1 Merging Trees

Input: L cohorts sets: $\{\{C_i^\ell\}_{i=1}^{n_\ell} \mid \ell = 1, \ldots, L\}$ and corresponding treatment effects and variances $\{\{(U(C),\ \sigma^2(C)) \mid C \in \{C_i^\ell\}_{i=1}^{n_\ell}\} \mid \ell = 1, \ldots, L\}$

Output: $\mathcal{S}_{out}$ and $\mathcal{T}_{out}$

Set $\mathcal{S}_{out} = \{C_i^1\}_{i=1}^{n_1}$ and $\mathcal{T}_{out} = \{(U_1(C),\ \sigma_1^2(C)) \mid C \in \mathcal{S}_{out}\}$
for $\ell = 2, \ldots, L$ **do**
    for $A \in \mathcal{S}_{out}$ **do**
        for $B \in \{C_i^\ell\}_{i=1}^{n_\ell}$ **do**
            $C = A \cap B$
            if $C \neq \emptyset$ **then**
                $\mathcal{S}_{out} = \mathcal{S}_{out} \cup \{C\}$
                $\mathcal{T}_{out} = \mathcal{T}_{out} \cup \{(U_m(C),\ \sigma_m^2(C)) \mid m = 1, \ldots, \ell\}$, where

$$U_m(C),\ \sigma_m^2(C) = \begin{cases} U_m(A),\ \sigma_m^2(A) & \text{for } m \leq l-1 \\ U_m(B),\ \sigma_m^2(B) & \text{for } m = \ell \end{cases}$$

            end if
        end for
    end for
    $\mathcal{S}_{out} = \mathcal{S}_{out} \setminus \{A\}$
    $\mathcal{T}_{out} = \mathcal{S}_{out} \setminus \{(U_m(A),\ \sigma_m^2(A)) \mid m = 1, \ldots, \ell - 1\}$
end for

Optimization Solution

Stochastic Optimization: the problem is stochastic since both the objective function and the constraints are not deterministic but are coming from a particular distribution (e.g., Gaussian).

$$\begin{aligned} \underset{\mathbf{x}}{\text{Maximize}} \quad & \mathbf{x}^T \mathbf{U}_0 \\ \text{subject to} \quad & \mathbf{x}^T \mathbf{U}_k \leq c_k \qquad \text{for } k = 1, \ldots, K. \\ & \sum_j x_{i,j} = 1 \ \ \forall i, \qquad 0 \leq \mathbf{x} \leq 1 \end{aligned}$$

$$\longrightarrow$$

$$\begin{aligned} \underset{\mathbf{x}}{\text{Maximize}} \quad & f(\mathbf{x}) = \mathbb{E}(\mathbf{x}^T \mathbf{U}_0) \\ \text{subject to} \quad & g_k(\mathbf{x}) := \mathbb{E}(\mathbf{x}^T \mathbf{U}_k - c_k) \leq 0, \qquad k = 1, \ldots, K. \\ & \textstyle\sum_j x_{ij} = 1 \ \ \forall\, i, \qquad 0 \leq \mathbf{x} \leq 1 \end{aligned}$$

Stochastic Approximation

Algorithm 2 Multiple Cooperative Stochastic Approximation

Input : Initial $\mathbf{x}_1 \in \mathcal{X}$, Tolerances $\{\eta_k\}_t, \{\gamma\}_t$, Iterations N
for $t = 1, \ldots, N$ **do**
    Estimate $\hat{G}_{k,t}$ for all $k \in 1, \ldots, K$ using (4).
    if $\hat{G}_{j,t} \leq \eta_{j,t}$ for all j **then**
        Set $h_t = F'(\mathbf{x}_t, \mathrm{U}_{0,t})$
    else
        Randomly select k^* from $\{k : \hat{G}_{k,t} > \eta_{k,t}\}$
        Set $h_t = G'_{k^*}(\mathbf{x}_t, \mathrm{U}_{k^*,t})$
    end if
    Compute $\mathbf{x}_{t+1} = P_{\mathbf{x}_t}(\gamma_t h_t)$
end for
Define $\mathcal{B} = \left\{1 \leq t \leq N : \hat{G}_{k,t} \leq \eta_{k,t} \ \forall k \in \{1, \ldots, K\}\right\}$
return $\hat{\mathbf{x}} := \frac{\sum_{t \in \mathcal{B}} \mathbf{x}_t \gamma_t}{\sum_{t \in \mathcal{B}} \gamma_t}$

Multiple Cooperative Stochastic Approximation At each step t it starts by estimating the constraint function.

$$\hat{G}_{k,t} = \frac{1}{L} \sum_{\ell=1}^{L} G_k(\mathbf{x}_t, \mathrm{U}_{k,\ell}).$$

- If feasible, chooses the gradient to be the gradient of the **objective**.

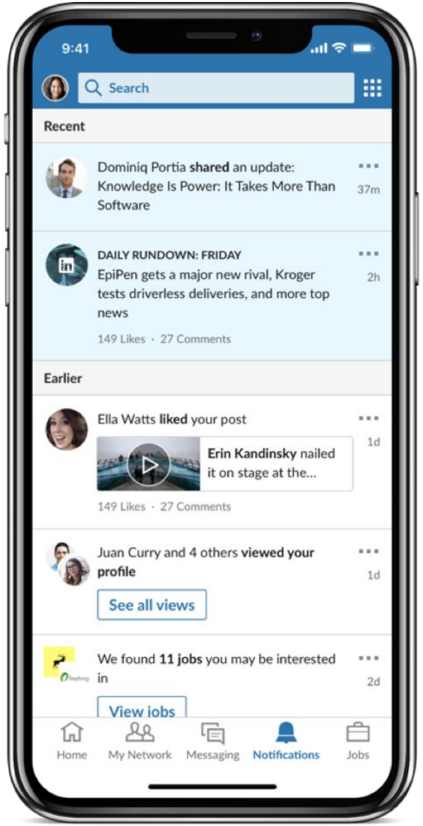
- Otherwise, from the set of violated constraints, it chooses a constraint **at random** and use the gradient of **that constraint**.

Overall Algorithm

- Run Randomized Experiments to collect data across various treatments and metrics
- Generate a cohort-level causal effects for the different parameters.
- Solve the Stochastic Optimization to generate the final treatment allocation to each cohort

Results

Notification Systems Results



- Metrics of Interest:

Metrics	Descriptions
Sessions (Objective)	Number of visits to the LinkedIn site/app
Notification Sends	Volume of notifications sent to members
Notification CTR	Click through rate on notifications
Total Disables	Number of total disables on notifications

Table 3: Metrics of Interest for Personalized Capping

Problem:

Maximize Sessions
S.t. Send $< s$
CTR $> c$
Disables $< d$

Notification System Results

- **Baseline:** Heuristic Cap A and B are based a cohort definition where members are grouped into four segments according to their visit frequency.

- **Personalized cap treatment** showed significant positive impact on Sessions, while the impact on the constraint metrics remained within acceptable bounds. It also outperforms the both heuristic solutions.

Metrics	ATE % Personalized Cap	ATE % Heuristic Cap A	ATE % Heuristic Cap B
Sessions	**+1.39%**	+1.31%	+0.54%
Notification Sends	**+1.64%**	+6.62%	+3.07%
Notification CTR	−1.24%	−1.73%	-1.18%
Total Disables	Neutral	+9.23%	Neutral

Table 4: Notification Cap Experiment Results

Discussions

Future work

A few non-trivial, but likely **impactful extensions** for future consideration include:

(1) Designing a more **cost-efficient data collection framework** or leveraging observational data to achieve the same performance would be beneficial.

(2) Users can potentially move in and out of cohorts. Extending this framework to incorporate the **dynamic nature of cohorts** could be an interesting research topic.

(3) Future work on generating one single optimal cohort definition based on effects from **multiple treatments with various metrics of interests** could further improve the method.

Reference

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

[2] Kinjal Basu and Preetam Nandy. 2019. Optimal convergence for stochastic optimization with multiple expectation constraints. arXiv preprint arXiv:1906.03401 (2019).

[3] Leo Breiman. 2001. Random Forests. Machine Learning 45, 1 (2001), 5–32.

[4] Oleg Sofrygin, Mark J. van der Laan, and Romain Neugebauer. 2017. simcausal R Package: Conducting Transparent and Reproducible Simulation Studies of Causal Effect Estimation with Complex Longitudinal Data. Journal of Statistical Software 81, 2 (2017), 1–47.

[5] Michał Sołtys, Szymon Jaroszewicz, and Piotr Rzepakowski. 2015. Ensemble methods for uplift modeling. Data Mining and Knowledge Discovery 29, 6 (2015), 1531–1559.

[6] Stefan Wager and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. J. Amer. Statist. Assoc. 113, 523 (2018), 1228–1242.

Thank you

Appendix

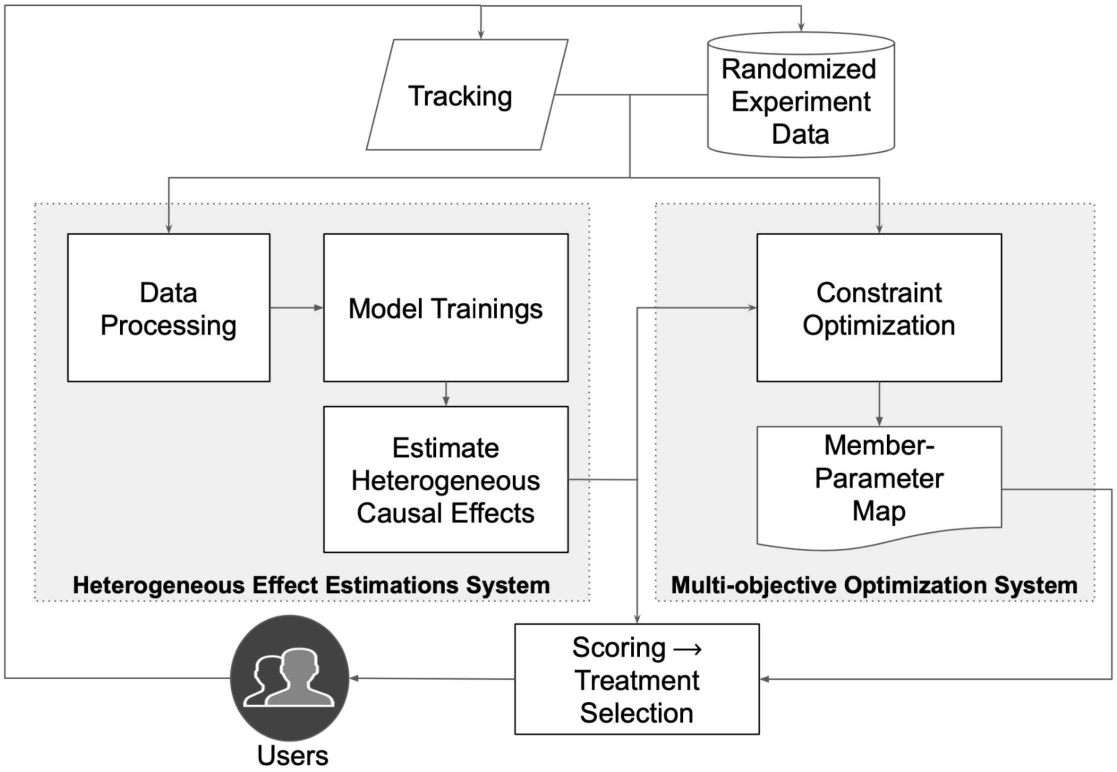
Member-level Heterogeneity

To estimate the heterogeneous causal effects at a **member level**, some of the options include:

(a) **Causal Forest**: The Causal Forest Algorithm is an extension of the Causal Tree which was inspired by Random Forest Algorithm and use ensemble learning to incorporate results from multiple tree models.

(b) **Two-Model Approach**: This is a baseline method (commonly applied in uplift modeling domain) that models the causal effect at a member level through the difference of the predicted response in the treatment and control models.

System Architecture



The general engineering architecture consists of two major components:

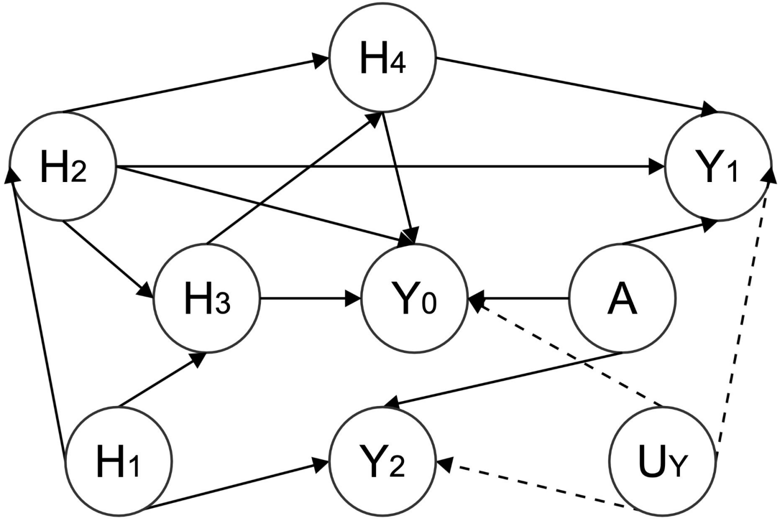
- One for **heterogeneous causal effect** estimations
- The other for the **optimization** module.

Simulation Analysis

We leverage simcausal R package [23] to generate simulation datasets under self-defined causal **Directed Acyclic Graphs** (DAG).

- *Aj* as the treatment variables
- *Yk* are the metrics (or response variables)
- *Uy* as a latent variable impacts *Yk*
- *Hm* as the heterogeneous variables

We simulate heterogeneity by introducing **interaction terms** between *Aj* and *Hm* on *Yk*.



Evaluation of Simulation

We consider the normalized mean of **individualized treatment effect** (ITE) for metric k at optimal x* as

$$\tau(\mathbf{x}^*)_k = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{3}(Y_{j,i,k} - Y_{0,i,k})\, z_{i,j}^*}{\mu_{0,k}},$$

($Y_{j,i,k} - Y_{0,i,k}$) represents the individualized treatment effect. We normalize the ITE by the control group mean μ to make results comparable across different simulated datasets.

Comparing all variants

(1) $HT.ST$: A heuristic cohort-level solution paired with stochastic optimization.

(2) $CT.ST$: Cohort-level estimations using Causal Tree model paired with stochastic optimization.

(3) $CF.DT$: Member-level estimations using the Causal Forest model [30] paired with deterministic optimization.

(4) $TM.DT$: Member-level estimations using a "Two-Model" approach (i.e., build two Random Forest [5] models) paired with deterministic optimization.

(5) *Global*: A best global allocation as baseline.

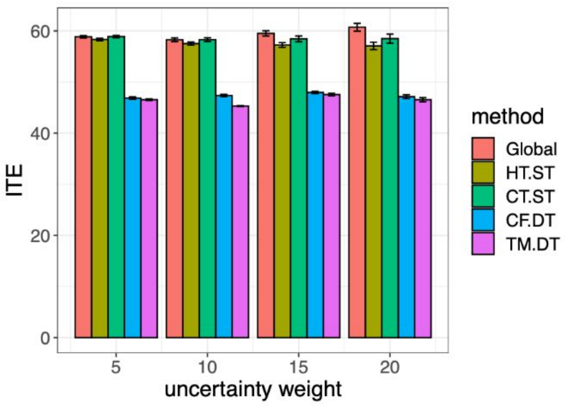
Analysis Results - Exist a global best

First scenario: Aligning the effect on the objective with that of the constraint metrics.

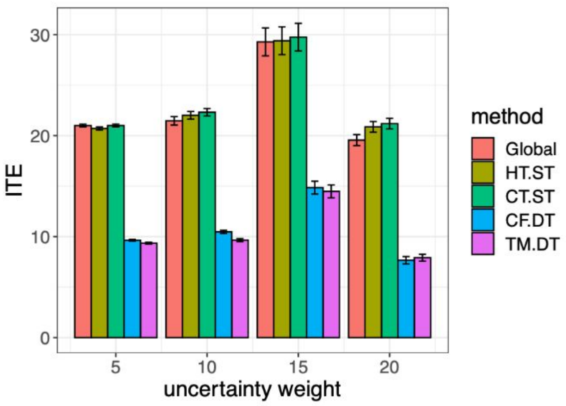
Benefit of the stochastic optimization:

- the cohort-level solutions paired with stochastic optimization ($HT.ST$ and $CT.ST$) perform almost at parity with the oracle global best solution $Global$.
- However, the member-level estimations paired with deterministic optimization ($CF.DT$ and $TM.DT$) show worse performance due to the high variance.

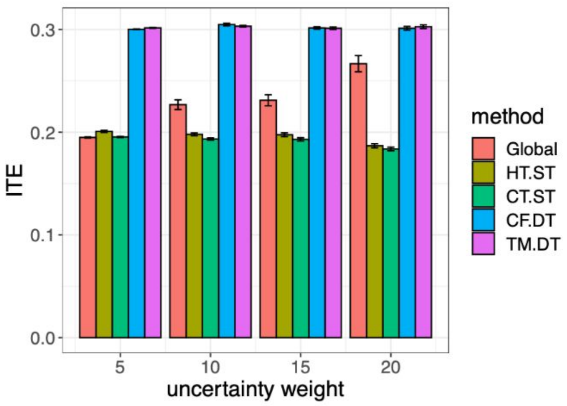
(a) Evaluation on the objective metric Y_0 if there exists a global best policy



(b) Evaluation on the constraint metric Y_1 if there exists a global best policy



(c) Evaluation on the constraint metric Y_2 if there exists a global best policy



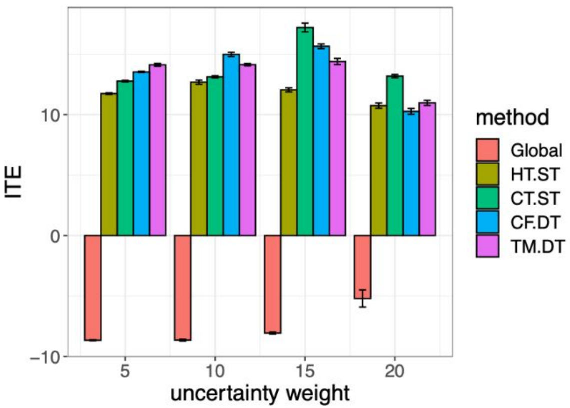
Analysis Results - No global best

Second scenario: the objective metrics move possibly in the opposite direction to some constraint metrics for some treatment.

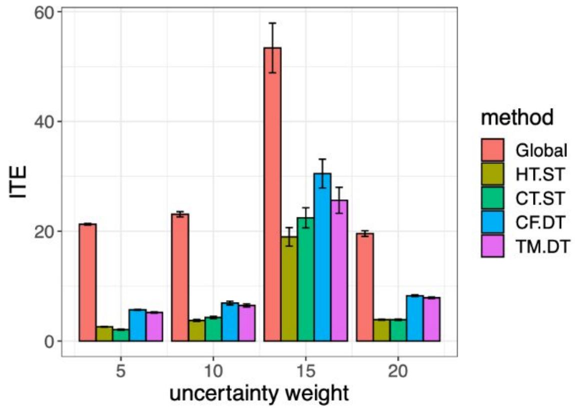
Benefit of heterogeneity estimation and personalization:

- All the proposed approaches perform better than the *Global* solution.
- With low noise levels, member-level solutions (*CF .DT* and *TM.DT*) perform better than the cohort-level solution (*HT.ST*, *CT.ST*). Along with an increase in the noise level, *CT.ST* quickly starts to catch up and can outperform the member-level solutions.

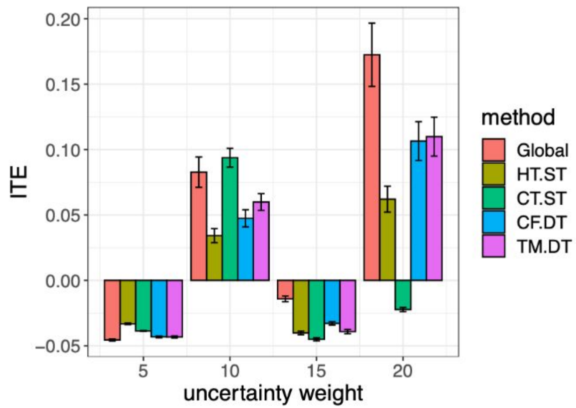
(d) Evaluation on the objective metric Y_0 if personalization can benefit the system



(e) Evaluation on the constraint metric Y_1 if personalization can benefit the system



(f) Evaluation on the constraint metric Y_2 if personalization can benefit the system



Reproducibility

We share example scripts for conduct **simulation analysis** in examining the proposed methods and stochastic optimization algorithms in the following Github link: https://github.com/tuye0305/prophet.