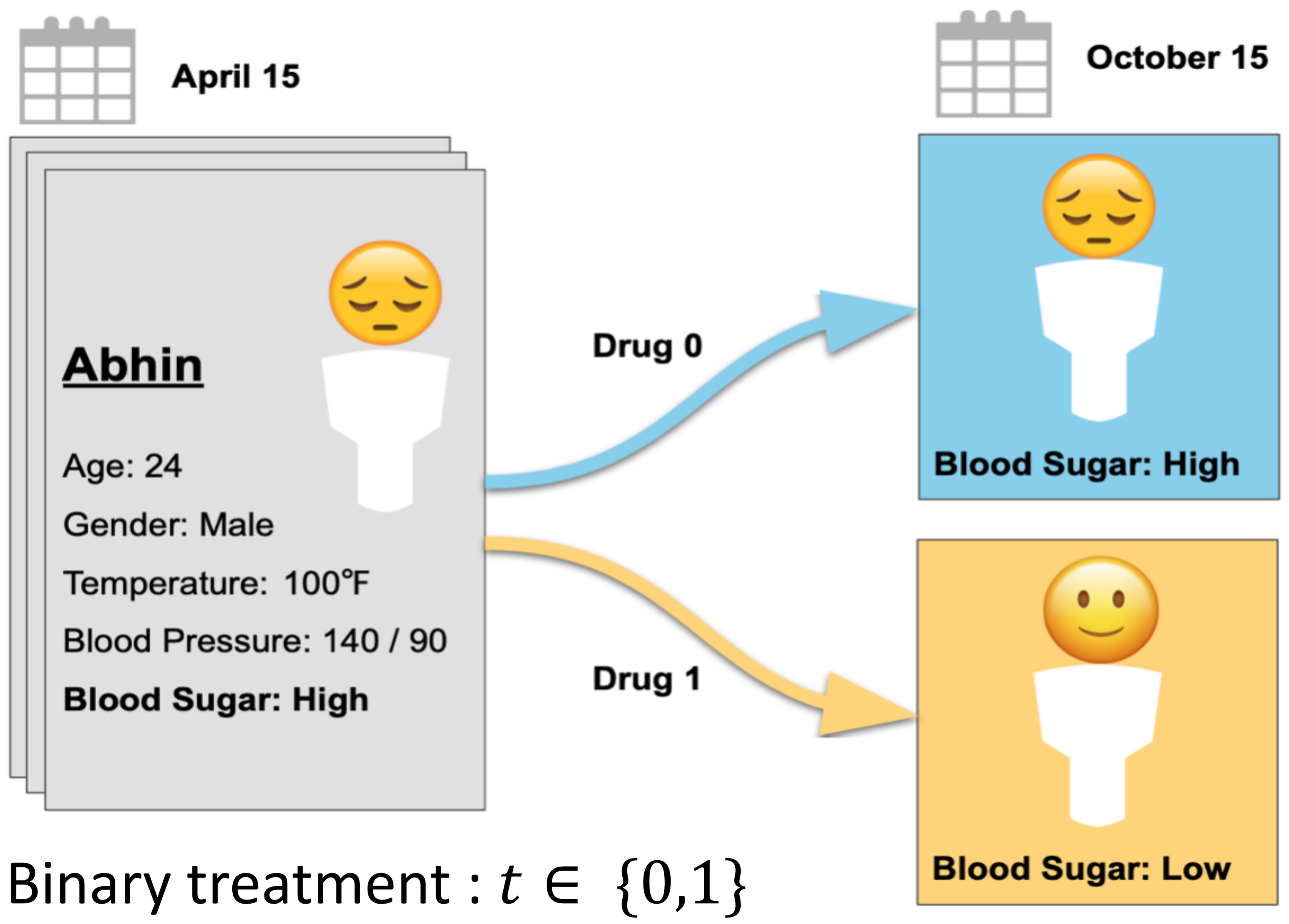
Treatment Effect Estimation Using Invariant Risk Minimization

Abhin Shah, Kartik Ahuja, Karthikeyan Shanmugam, Dennis Wei, Kush R. Varshney, Amit Dhurandhar

Paper Number 4195

Individual Treatment Effect

Goal- Understand the causal effect of a treatment t on an individual with features x from observational data



- Binary treatment : $t \in \{0,1\}$
- Potential outcomes : y_i for $t = i$
- Observational data : Observe $y_f = (1-t) * y_0 + t * y_1$

$$\boxed{ITE(x) = y_1(x) - y_0(x)}$$

- Challenge – Treatment assignment bias

Common regression adjustment methods

- T-learner – use separate base-learners to estimate the outcome under control and the outcome under treatment
- S-learner – use one base-learner to estimate the outcome using the features and the treatment assignment

This work

Objective **–** Build methods for robust ITE estimation

Key Idea **–** IRM can be used to tackle treatment assignment biases in ITE settings

Invariant Risk Minimization

Goal- Identify which properties of the training data describe spurious correlations and which properties represent phenomenon of interest for out-of-distribution generalization

Train distribution $\neq$ Test distribution

Correlation versus causation

- ERM picks spurious correlation i.e., the background
- IRM focuses on causative features i.e., cow's shape

Requirement – Training data from distinct environments

Proposed methods

- IRM_1 – S-learner with IRM for square loss and linear function class as the base-learner.
- IRM_2 – T-learner with IRM for square loss and linear function class as the base-learners.

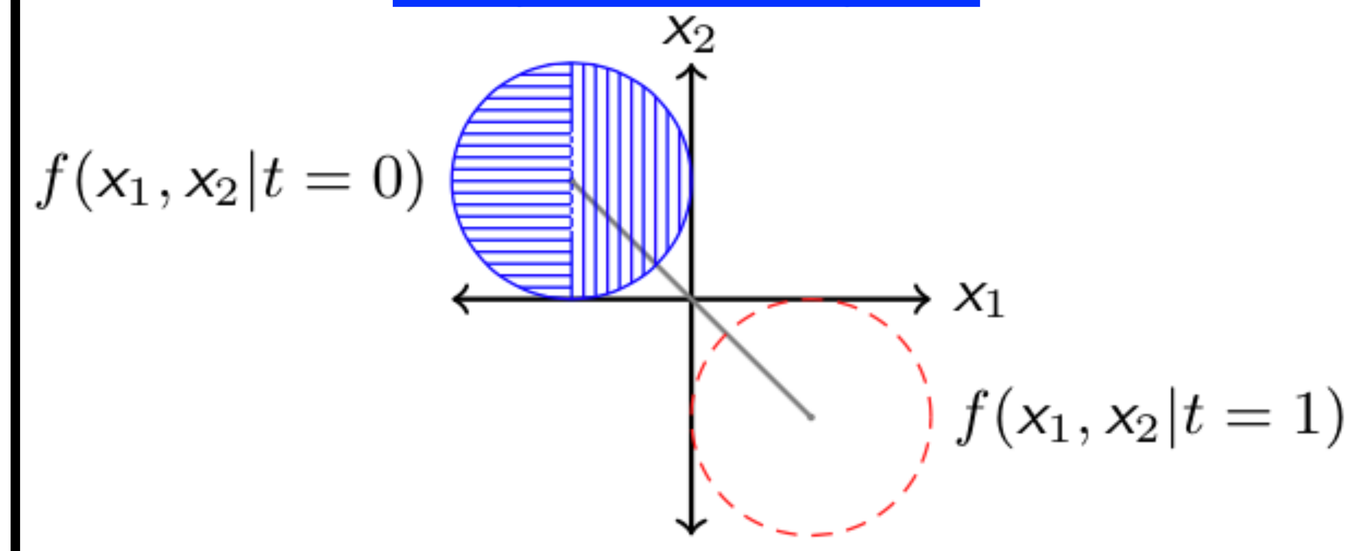
Baselines

- OLS/LR1 – S-learner with ERM for square loss and linear function class as the base-learner.
- OLS/LR2 – T-learner with ERM for square loss and linear function class as the base-learners.

Procedure

1. Observational data $D = \{(x^{(i)}, t^{(i)}, y_f^{(i)}) : i = 1, \cdots, n\}$
2. Create n_e domains from $D = \{D_j : j = 1, \cdots, n_e\}$
3. Apply ERM on D and IRM on $(D_j : j = 1, \cdots, n_e)$

Toy example



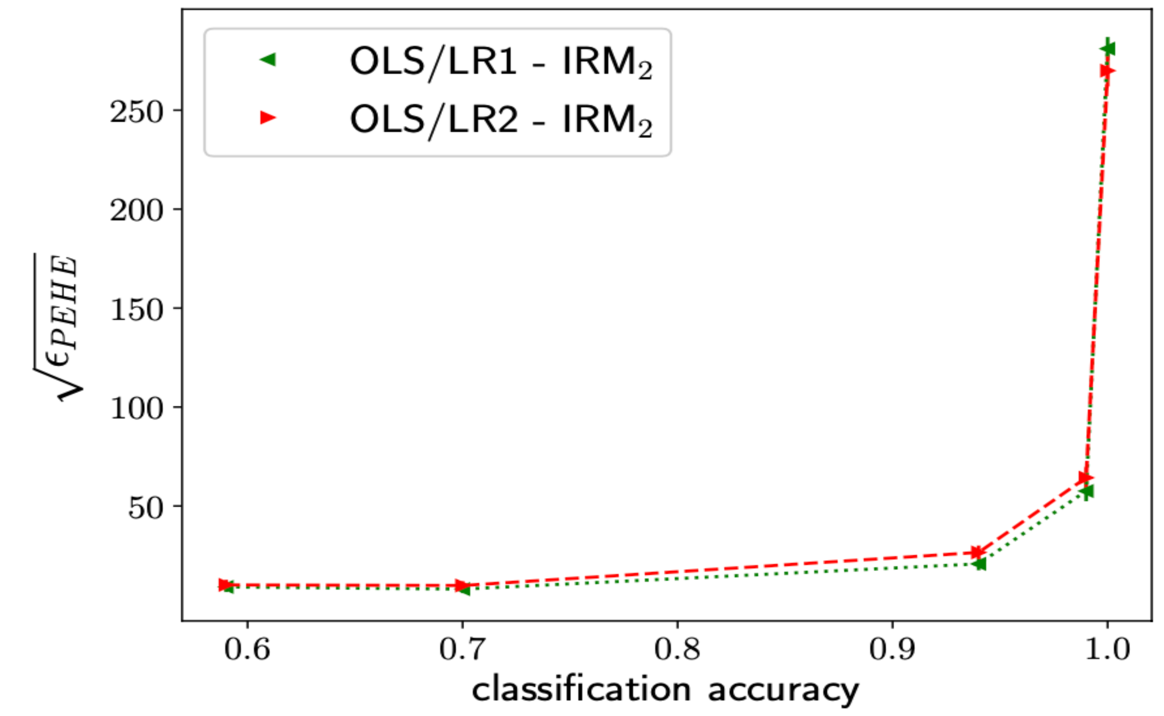
Experiments

- $t \sim Bernoulli(0.5)$
- $x \mid t \sim \mathcal{N}(\mu_t, \Sigma)$
- $y_t | x, t \sim \mathcal{N}(xTAtx + x^T b_t + ct, \sigma^2)$
- $e \sim Bernoulli(0.5)$

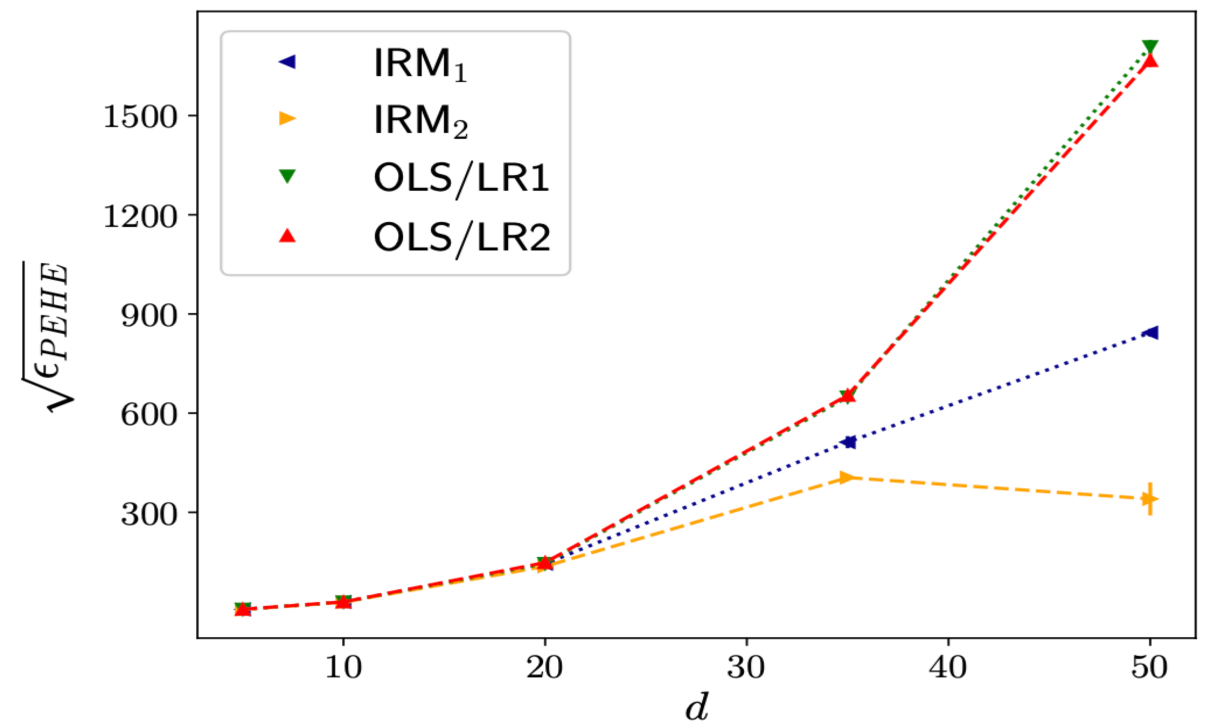
Performance metrics

- $\epsilon_{PEHE} = \frac{1}{n} \sum_{i=1}^{n} (ITE(x_i) - \widehat{ITE}(x_i))^2$

Plots



1. $\sqrt{\{\epsilon_{PEHE}\}}$ difference vs treatment group classification accuracy



2. $\sqrt{\{\epsilon_{PEHE}\}}$ vs dimensions of features