ABC3: Active Bayesian Causal Inference with Cohn Criteria in Randomized Experiments

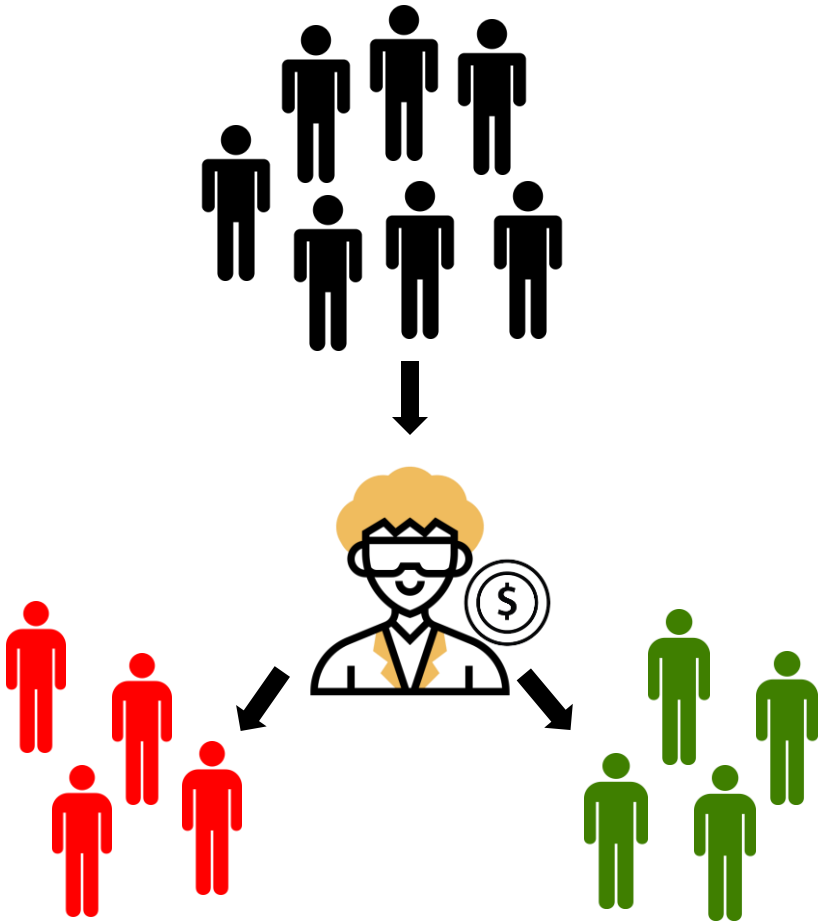
April 4, 2025

Taehun Cha
Ph.D. Candidate
Korea University

Contents

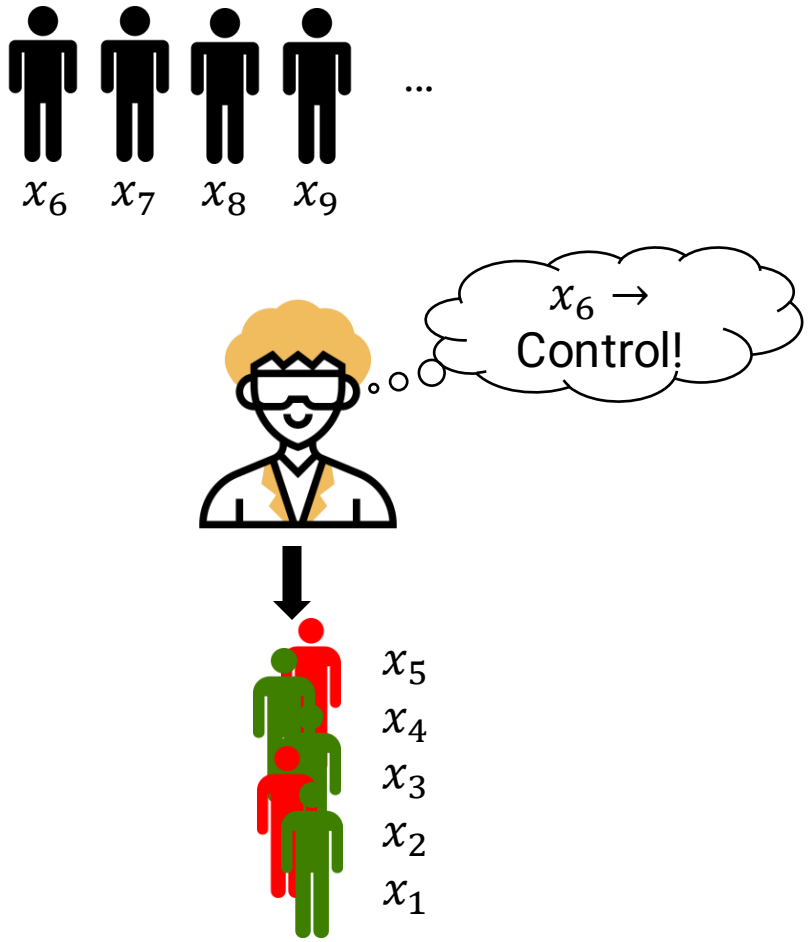
- Active Learning in Randomized Experiment
- Algorithm: ABC3
- Theory: MMD and Type 1 Error
- Discussion and Conclusion

Active Learning in Randomized Experiment



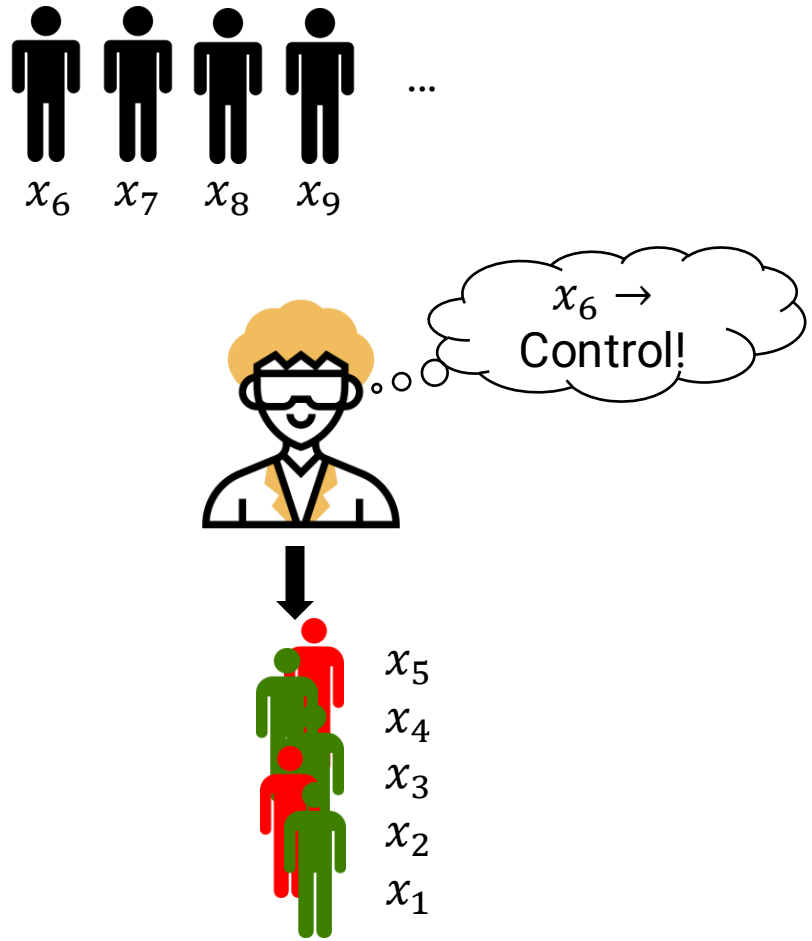
- Randomized experiment divides subjects into two groups, treatment and control groups.
- However, randomized experiment is usually expensive, so an efficient experiment design is desirable.

Active Learning in Randomized Experiment



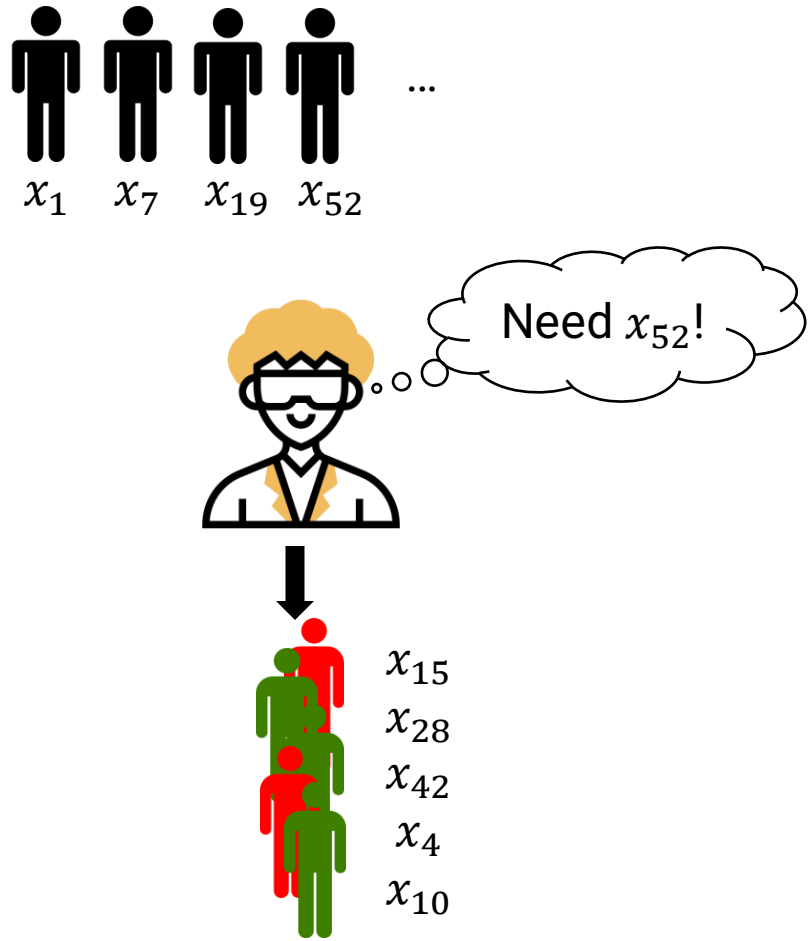
- Several works suggested efficient treatment allocation design given outcome variance, subject covariate, etc.
- However, these works assume the experimental subjects are given, not actively choosable.

Active Learning in Randomized Experiment



- What if we already know the covariate information in prior?
  - Internet companies may know personal information before A/B test.
  - Pharmaceuticals have experiment applicants' personal information.
- Can we rationalize randomized experiment by choosing 'proper' subject at time t?

Active Learning in Randomized Experiment



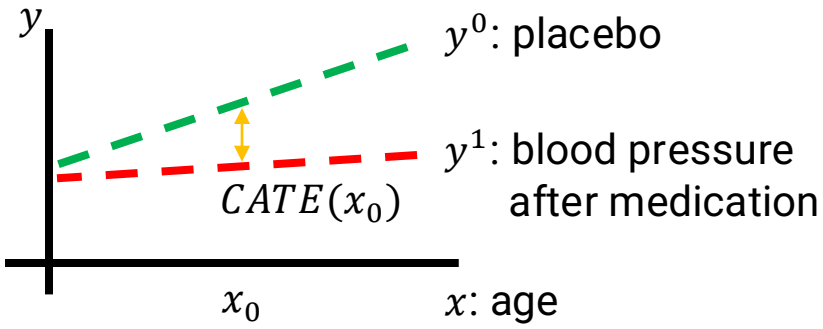
- Active learning: a practitioner sequentially choose unlabeled data points and ask an oracle to label them.
- Algorithm
  - Goal: Estimate $f(x)$ w.r.t. utility function u_t
  - For $t \in 1, \ldots, T$:
    - Compute utility function $u_t(x)$, x: unseen data point
    - Choose $x_t = x^*$ which optimizes u_t
    - Obtain y_t and train $\hat{f}$ with $\{(x_i, y_i)\}_{i=1}^{t}$
- Goal: Make randomized experiment efficient with active learning framework

Algorithm: ABC3

- Conditional Average Treatment Effect ($f(x)$)

$$CATE(x) = E[Y^1 - Y^0 | X = x]$$

where Y^a: potential outcome, X: covariate.



- Expected precision in estimation of heterogeneous effect

$$\epsilon_{PEHE}(\widehat{CATE}_t) = \int_X \left(\widehat{CATE}_t(x) - CATE(x)\right)^2 dP(x)$$

where $\widehat{CATE}_t = \hat{y}_t^1 - \hat{y}_t^0, \hat{y}_t^a$: a regressor at t.

- To optimize PEHE, we should assume the existence of population parameter, CATE.

Algorithm: ABC3

- We utilize Bayesian framework by defining

$$\widehat{CATE}_{\Omega}(x) = \hat{y}_{\Omega}^{1} - \hat{y}_{\Omega}^{0}$$

$$\epsilon_{PEHE}^{\Omega}(\widehat{CATE}_t) = \int_X \left(\widehat{CATE}_t(x) - \widehat{CATE}_{\Omega}(x)\right)^2 dP(x)$$

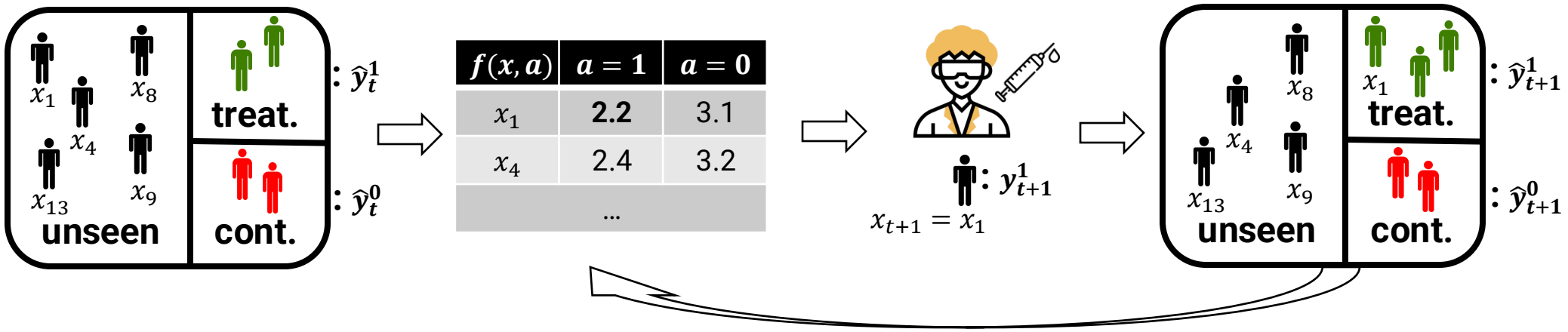
  where $\hat{y}_{\Omega}^{a}$: a regressor trained with whole oracle data set.
- Define a utility function $u_t(x) = E_{t+1}\left[\epsilon_{PEHE}^{\Omega}\left(\widehat{CATE}_{t+1}\right)\right]$, assuming $x_{t+1} = x$.
- **(Theorem)** Assume $Y^a \sim GP\left(0, k(x, x')\right), \hat{y}_t^a(x) = E_t[Y^a(x)]$. Then under mild assumptions,

$$argmin_{x_{t+1}, a_{t+1}} E_{t+1}\left[\epsilon_{PEHE}^{\Omega}\left(\widehat{CATE}_{t+1}\right)\right] =$$

$$argmin_{x_{t+1}, a_{t+1}} \int_X V_{t+1}[Y^1(x)] + V_{t+1}[Y^0(x)] dP(x)$$

(Roughly, minimizing future error ⇔ minimizing future variance)

Algorithm: ABC3

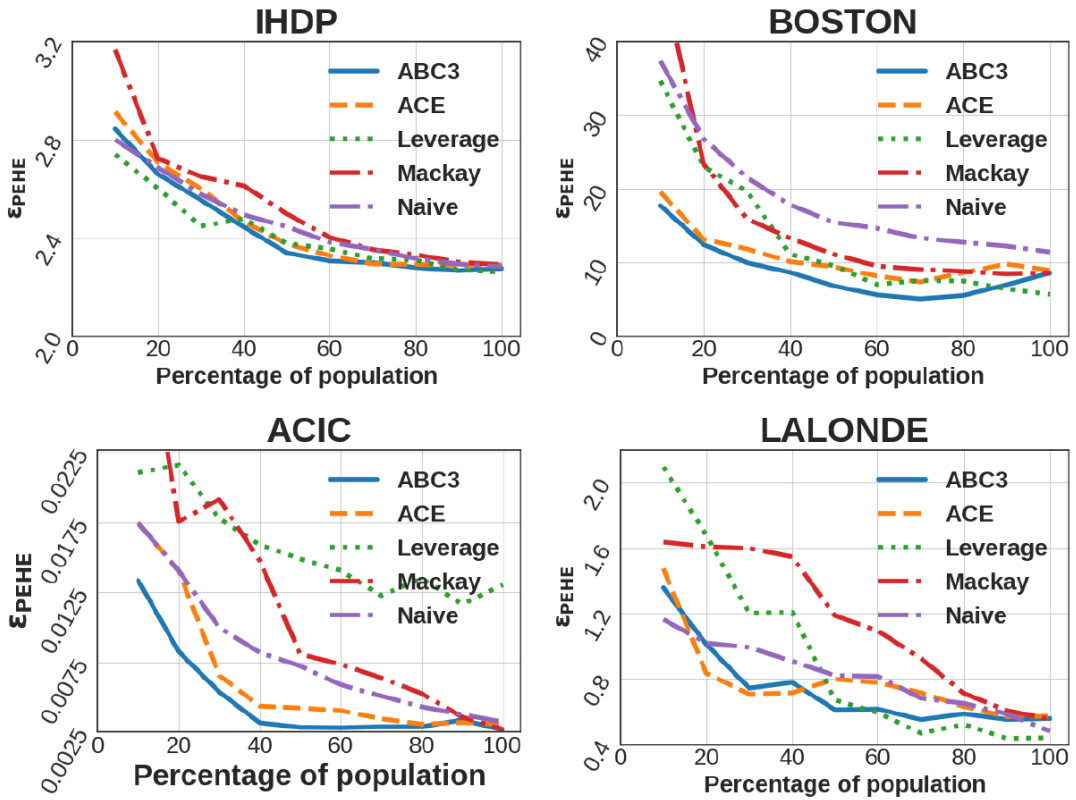


- Algorithm
  - Compute $f(x,a) = \int_X V_{t+1}[Y^1(x)] + V_{t+1}[Y^0(x)]dP(x)$ when $x_{t+1} = x \in$ unseen points.
  - Choose x_{t+1} and a_{t+1} which minimizes $f(x,a)$ and observe y_{t+1}^a.
  - Train a Gaussian process $\hat{y}_{t+1}^a$ with the new data point.
- We propose an efficient way to compute the quantity without computing the inverse of the kernel matrix for every x. (Proposition 4.1.)

Algorithm: ABC3

- Cohn (1994) suggested a similar active learning criteria which minimizes integrated predictive variance.
- The future variance assuming observation on x_{t+1}, $V_{t+1}[Y^a(x)]$, does not depend on outcome y_{t+1} (computable in prior).
- Experiment
  - Naïve: Usual randomized experiment
  - Mackay: choose a point of maximum variance
  - ACE: minimizing the predictive variance while accessing test data set
  - Leverage: choose subsamples which are optimal under linearity assumption

Algorithm: ABC3

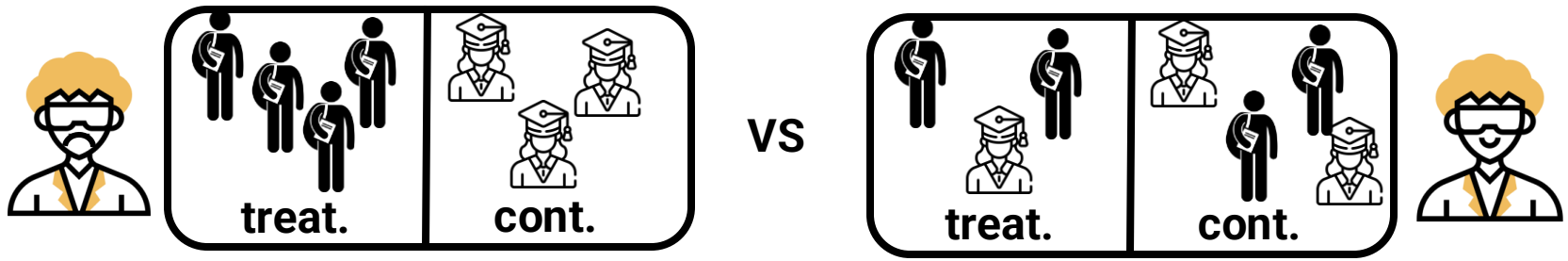


- Boston: ABC3 at 20% > Naïve at 100%
- ACIC: ABC3 at 40% > Naïve at 100%
- Mackay underperforms even Naive policy most times
- Leverage significantly underperforms other policies in ACIC

  -> vulnerability of linearity assumption
- ABC3 outperforms ACE without access to the test data set

Theory: MMD

- Balance between the treatment/control groups is crucial.
  - Consider a study on the causal effect of online lectures where treatment group is undergraduate students while control group is graduate students.



- Balance measure: Maximum Mean Discrepancy

$$MMD(P, Q, F) = \sup_{f \in F} E_{x \sim P(x)}[f(x)] - E_{y \sim Q(y)}[f(y)] = \left|\left|\mu_P - \mu_Q\right|\right|^2$$

where μ_P: mean embedding of P in RKHS.

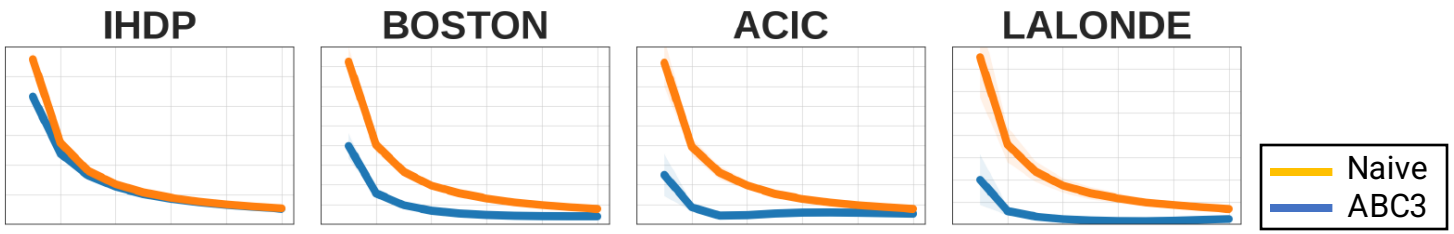
- MMD measures the difference in mean embeddings of two groups, P and Q.

Theory: MMD

- **(Theorem)** Let P_t^a: treatment/control group covariate distribution at t, I_t^a: index set of each group at t, and λ^*: maximum eigenvalue of the kernel matrix of whole covariates. Then under some mild condition,

$$MMD(P_t^1, P_t^0, F)^2 \leq 4\left(\frac{\lambda^*}{|I_t^1|} + \frac{\lambda^*}{|I_t^0|}\right) + 2\int_X V_t[Y^1(x)] + V_t[Y^0(x)]dP(x)$$

where the first term is minimized as time proceeds, and the second term is ABC3's optimization target.



(Roughly, ABC3 achieves a balance between treatment and control groups.)

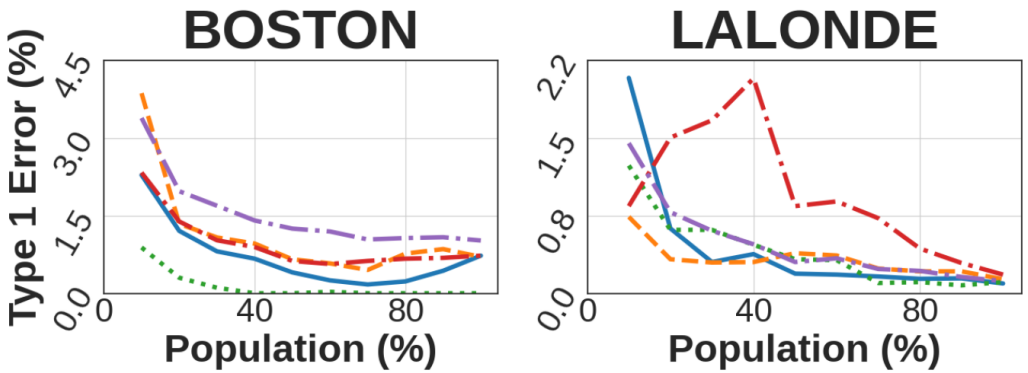
Theory: Type 1 Error

- Type 1 Error rate: If null hypothesis $H^0: Y^1 = Y^1$ holds,

$$P_t[Type\ 1\ Error(x)] = P_t[|Y^1(x) - Y^0(x)| > \alpha]$$

where α: decision threshold.

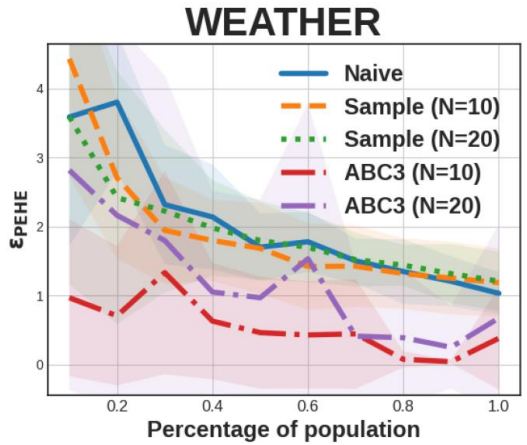
- **(Theorem)** Under Fisher's Sharp Null (i.e. $CATE(x) = 0$), ABC3 minimizes the upper bound of $\int_X P_{t+1}[Type\ 1\ Error(x)]\, dP(x)$



(Roughly, ABC3 minimizes Type 1 error when there is no treatment effect.)

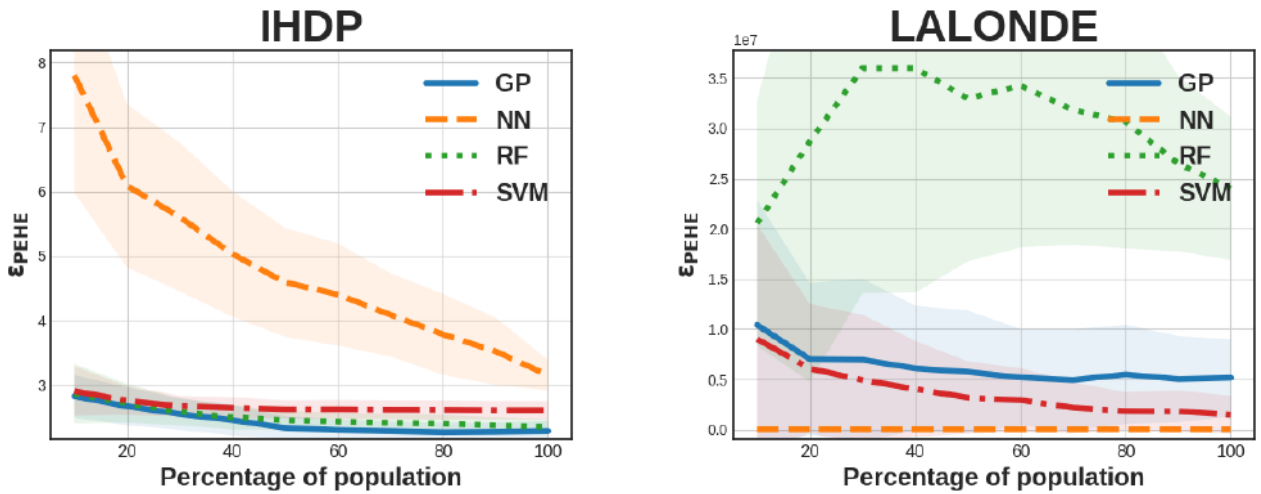
Discussion

- Extending ABC3 to large Weather data set
  - **Sample** from (un)observed covariates and compute kernel matrix between them.
  - **Optimize** to find x^* which minimizes our target quantity.
  - Then choose x_{t+1} near x^*.



Discussion

- Plugging-in Other Regressor
  - Use GP-based ABC3 only for sampling
  - Then use another type of models for regressor $\hat{y}^a$
  - Performance depends on data sets



Conclusion

- We introduce ABC3, active learning algorithm for causal inference from Bayesian perspective.
- ABC3 outperforms other algorithms by balancing treatment/control groups and minimizes type 1 error rate.

Paper

Code

Personal