A Novel Evaluation Methodology for Assessing Off-Policy Learning Methods in Contextual Bandits

Negar Hassanpour* and Russell Greiner*

*Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

1. Introduction

Goal: Finding the best personalized treatment as suggested by a policy **π(t|x)**, derived by an off-policy learning method trained on contextual bandit data

Requires: A way to effectively evaluate any proposed policy **π**, perhaps by running it on a range of **realistic*** bandit datasets **{ [x_i, t_i, y_i] }** exhibiting various levels/types of sample selection bias

***realistic**: similar to a real contextual bandit dataset -- especially **X**

> Standard supervised machine learning data is problematic **(1)**

> Better to use **X** from a real bandit data [here, a Randomized Control Trial (RCT)]

Background: Existing approach for synthesizing a bandit dataset

Overview:

- **Input:** a multi-binary-label supervised dataset $\mathcal{D}^* = \{\, [\,(x_i,\ t_i^*)\,]\,\}_{i=1..n}$ with $t_i^* \in \{0,1\}^k$
- **Output:** a bandit dataset
  - **outcome (Y):** Jaccard index calculated from t and t*
  - **h (t|x) :** the logging policy; a function [usually logistic regression (LR)]; parameters learned from a small portion of training set

Shortcomings: in addition to **(1)** above …

2. Not clear how to map a **binary multi-label** to a medical treatment (e.g., problem of drug interactions)
3. Jaccard index as outcome implies assigning **equal importance** to various treatment options
4. Assumes **known underlying mechanism** of treatment selection (e.g., in ad-placement); i.e., **h (t|x)** is known

2. Proposed Approach

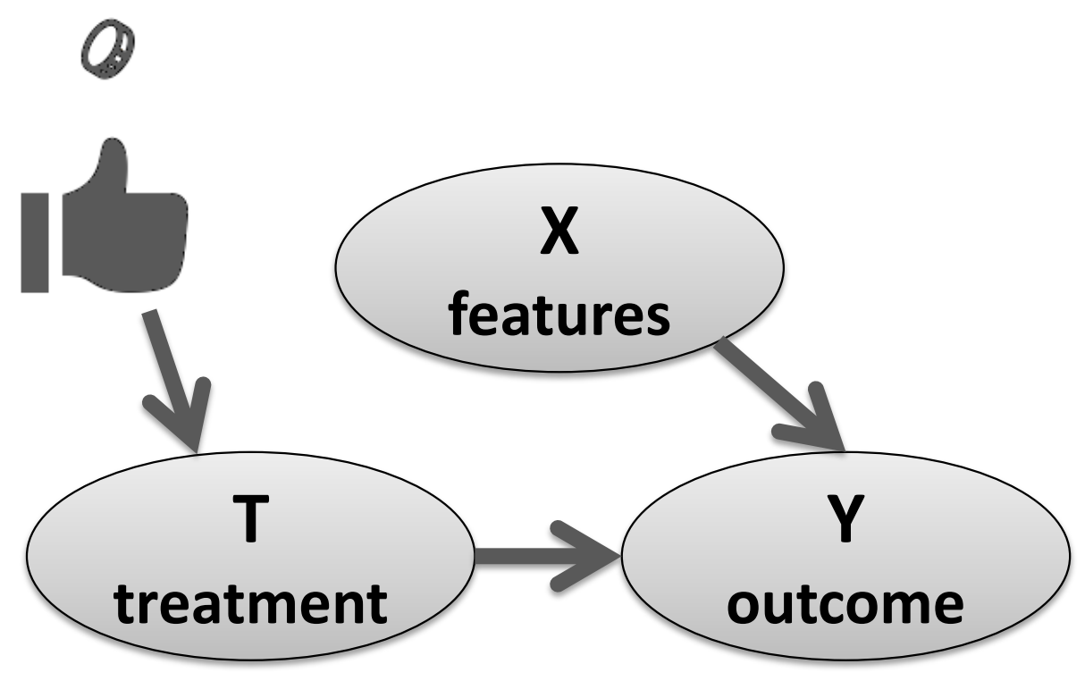
Overview:

- **Input:** a Randomized Control Trial (RCT) dataset **{ [x_i, t_i, y_i] }** with two treatment arms: RCT*
- **Output:** a new bandit dataset with same **X**, …
  - **outcome (Y):** fit a Gaussian Process to each treatment arm { **$\mu_t(x)$** and **$\sigma_t(x)$** }; define counterfactual outcomes $\hat{y}_t(x_i) = \mu_t(x_i) + k_t \times \sigma_t(x_i)$ such that statistics from any sampled RCT match RCT* (see 2.A)
  - **$h_{\alpha,\beta}$(t|x) :** a sigmoid function with varying parameters to generate different bandit datasets (see 2.B)

Pipeline:

In-put

Original Randomized Control Trial (RCT*)



X	x_1		x_2		x_3		
T	0	1	0	1	0	1	…
Y	$y_0(x_1)$	?	?	$y_1(x_2)$	?	$y_1(x_3)$	

Calculate RCT*'s Statistical Measures (2.A)

- Average Treatment Effect

$$ATE = \frac{1}{N_1}\sum_{t_i=1} y_1(x_i) - \frac{1}{N_0}\sum_{t_j=0} y_0(x_j)$$

- Coefficients of Determination

$$R_t^2 = 1 - \left[\sum_{t_i=t} y_t(x_i) - \hat{y}_t(x_i)\right]^2 \Bigg/ \left[\sum_{t_i=t} y_t(x_i) - \bar{y}\right]^2 \quad for\ \ t \in \{0,1\}$$

Derive Counterfactual Functions (2.A)

- Fit Gaussian Process to each treatment arm to get $\{\mu_t(.),\ \sigma_t(.)\}$
- Set counterfactual functions to: $\boxed{\hat{y}_t(x_i) = \mu_t(x_i) + k_t \times \sigma_t(x_i)}$
- Find k_t such that any new sampled RCT (i.e., synthetic) has statistics similar to RCT*

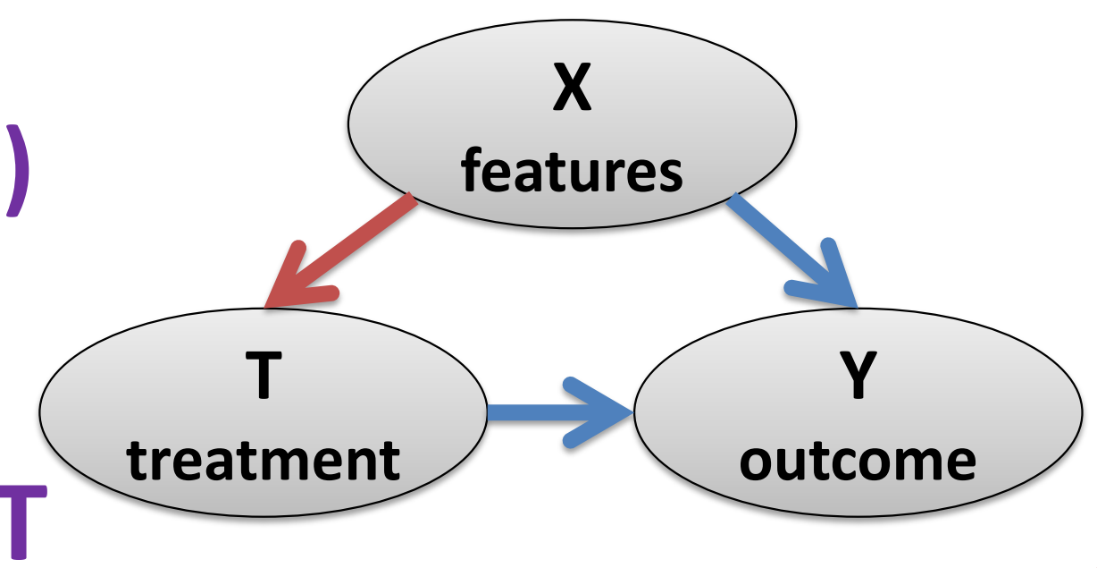
Compute Counterfactual Outcomes (2.A)

- Fill in the data table with counterfactual outcomes

X	x_1		x_2		x_3		
T	0	1	0	1	0	1	…
Y	$y_0(x_1)$	$\hat{y}_1(x_1)$	$\hat{y}_0(x_2)$	$y_1(x_2)$	$\hat{y}_0(x_3)$	$y_1(x_3)$	

Decide the bias level and tune the logging policy (2.B)

- Vary **α** and **β** parameters of the sampling function **$h_{\alpha,\beta}$(t|x)** to control the amount/type of the sample selection bias induced by the link from **X** to **T**



Out-put

Synthesize many Observational Studies (2.B)

Outcomes for Observational Study #1

Outcomes for Observational Study #2

X	x_1		x_2		x_3		
T	0	1	0	1	0	1	
Y_1	?	$\hat{y}_1(x_1)$	$\hat{y}_0(x_2)$	?	?	$y_1(x_3)$	…
Y_2	$y_0(x_1)$	?	$\hat{y}_0(x_2)$	?	$\hat{y}_0(x_3)$	?	
⋮	…						

2.A Defining outcomes: Preserve characteristics of the RCT*

Goal: A synthetic bandit dataset as similar as possible to a real bandit dataset

→ Attempt to preserve the following statistical characteristics of RCT*

- Average Treatment Effect (**ATE**)
- Coefficient of Determination (**R_0^2** and **R_1^2**)

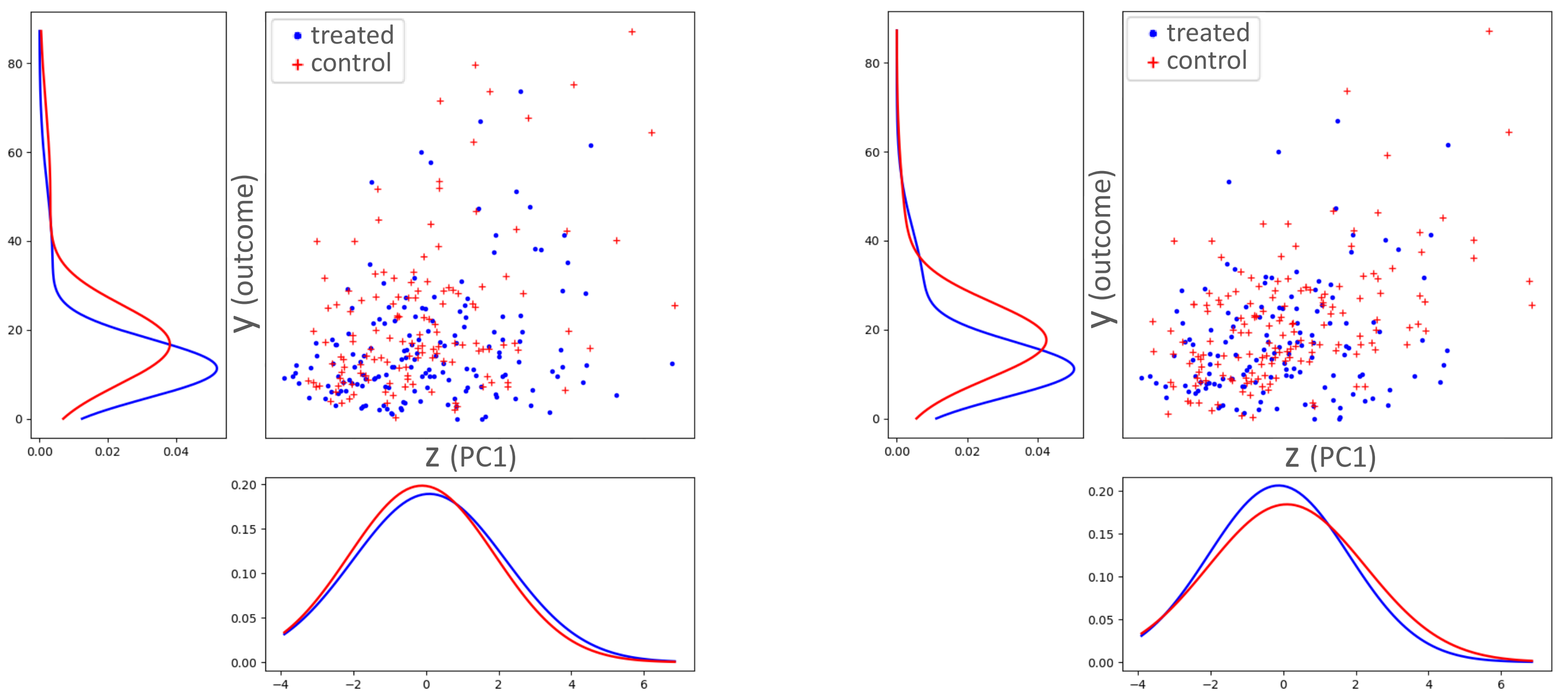
Counterfactual for a **t_i** = 1 subject with observed outcome **$y_1(x_i)$** is: $\hat{y}_0(x_i) = \mu_0(x_i) + k_0 \times \sigma_0(x_i)$

k_0 is determined such that the average personalized treatment effect calculated on the **N_1** subjects with **t** = 1, i.e.,

$$A\hat{T}E_1 = \frac{1}{N_1}\sum_{t_i=1}\left(y_1(x_i) - \hat{y}_0(x_i)\right)$$

matches the **ATE** calculated on the RCT*

$$k_t = \left(ATE - (2t-1)\frac{1}{N_{\neg t}}\sum_{t_i=\neg t}\left(\mu_t(x_i) - y_{\neg t_i}(x_i)\right)\right) \Bigg/ \frac{1}{N_t}\sum_{t_j=t}\sigma_t(x_j)$$



(a) RCT*

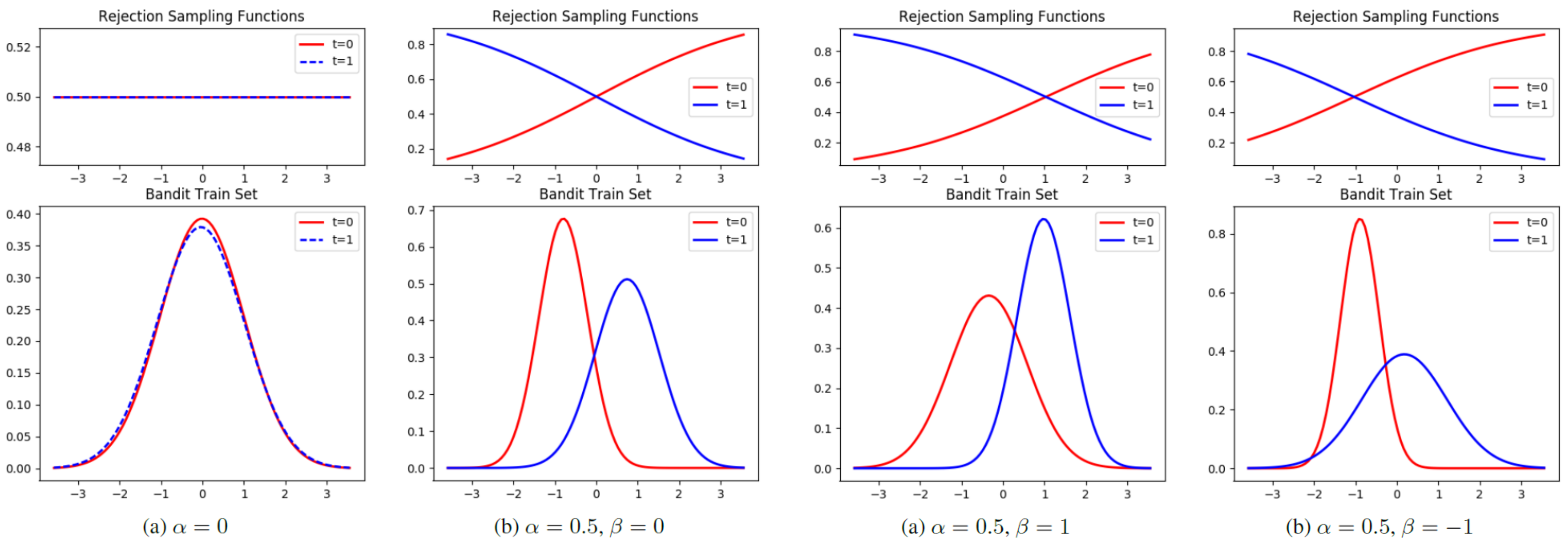
(b) Synthetic RCT

2.B Various logging policies

Logging policy **$h_{\alpha,\beta}$(t|x)** is a sigmoid with **α** and **β** parameters; **z=z(x)** is function of **x** that induces the sample selection bias. **z** could be:

- any covariate such as age, gender, etc. OR principal component #1, etc.
- amount of improvement in outcome if the best treatment was received

$$h_{\alpha,\beta}(t=1\,|\,x) = \frac{1}{1+e^{-\alpha(z(x)-\beta)}}$$



(a) $\alpha = 0$ (b) $\alpha = 0.5, \beta = 0$

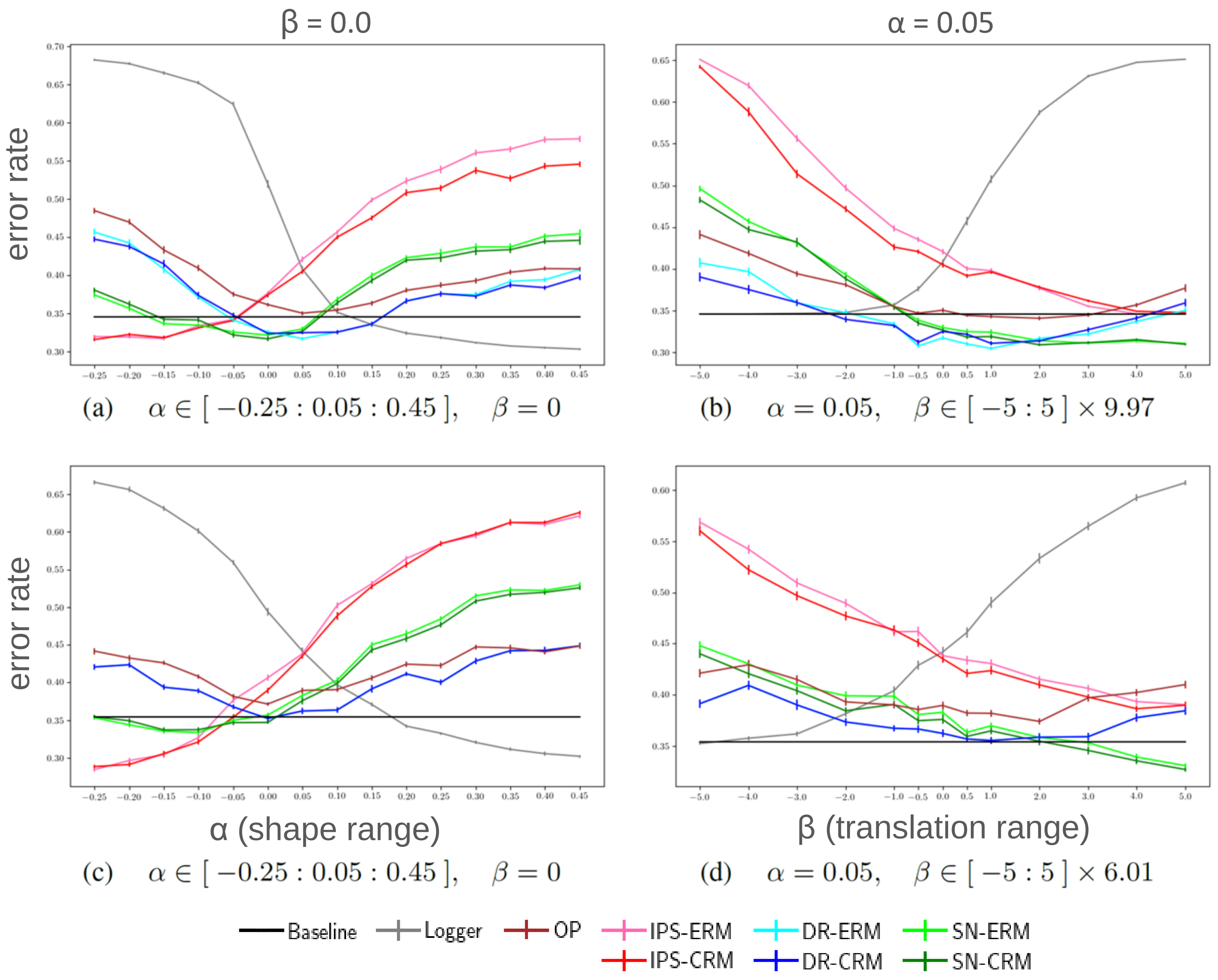
Effect of changing α on selection bias; note α=0 generates an RCT

(a) $\alpha = 0.5, \beta = 1$ (b) $\alpha = 0.5, \beta = -1$

Effect of changing β on conservatism towards preferring one treatment to the other

3. Experimental Results

Evaluated several prominent off-policy learning methods in contextual bandits: Outcome Prediction (**OP**), Inverse Propensity Scoring (**IPS**), Doubly Robust (**DR**), and Self-Normalized (**SN**) with either Empirical or Counterfactual Risk Minimization principle (**ERM** and **CRM** respectively), on several contextual bandit datasets exhibiting various levels of sample selection bias (**α** and **β**), synthesized using two RCT* datasets.



(a) $\alpha \in [\,-0.25 : 0.05 : 0.45\,], \quad \beta = 0$

(b) $\alpha = 0.05, \quad \beta \in [\,-5 : 5\,] \times 9.97$

(c) $\alpha \in [\,-0.25 : 0.05 : 0.45\,], \quad \beta = 0$

(d) $\alpha = 0.05, \quad \beta \in [\,-5 : 5\,] \times 6.01$

Mean ± 0.1×SDev of the classification error rates on Acupuncture (top) and Hypericum (bottom) datasets

Analyses show that different off-policy learning methods exhibit different competencies under various conditions of sample selection bias. The proposed evaluation framework enables us to tease out these differences and select the appropriate method for each real-world application.