Learning representations for counterfactual inference

Fredrik D. Johansson*[2], Uri Shalit*[1], David Sontag[1]

*Equal Contribution

NIPS 2016 Deep Learning Symposium

December 2016

1 NYU

2 CHALMERS UNIVERSITY OF TECHNOLOGY

Talk today about two papers

- Fredrik D. Johansson, Uri Shalit, David Sontag
  "*Learning Representations for Counterfactual Inference*"
  ICML 2016
- Uri Shalit, Fredrik D. Johansson, David Sontag
  "*Estimating individual treatment effect: generalization bounds and algorithms*"
  arXiv:1606.03976

Code: https://github.com/clinicalml/cfrnet

Causal inference from observational data

- Patient “Anna” comes in with hypertension
  - Asian, 54, history of diabetes, blood pressure 150/95, …
- Which of the treatments t will cause Anna to have lower blood pressure?
  - Calcium channel blocker ($t = 1$)
  - ACE inhibitor ($t = 0$)
- Dataset of ***observational data*** from many patients: medications, blood tests, past diagnoses, demographics …

Causal inference from observational data

- Patient “Anna” comes in with hypertension
  - Asian, 54, ... 50/95, ...
- Which of t... o have lower blood pres...
  - Calcium c...
  - ACE inhib...

How to best use *observational data* for individual-level causal inference?

- Dataset of ***observational data*** from many patients: medications, blood tests, past diagnoses, demographics …

Causal inference from observational data: Job training

- 1,000 unemployed persons
- Job training program with capacity of 100
  - Training ($t = 1$)
  - No training ($t = 0$)
- Who should get job training?
  - For which persons will job training have the most impact?
- Observational data about thousands of people:
  job history, job training, education, skills, demographics…

Observational data

- **Dataset of features, actions and outcomes**
- **We do not control the actions**
- **We do not know the model generating the actions**


Causal inference from observational data and reinforcement learning

- Robot on the sideline, learning by observing other robots playing robot football
- Sideline-robot does not know the playing-robots' internal model
- Form of off-policy learning, learning from demonstration

Outline

Background

Model

Experiments

Theory

Outline

Background

Model

Experiments

Theory

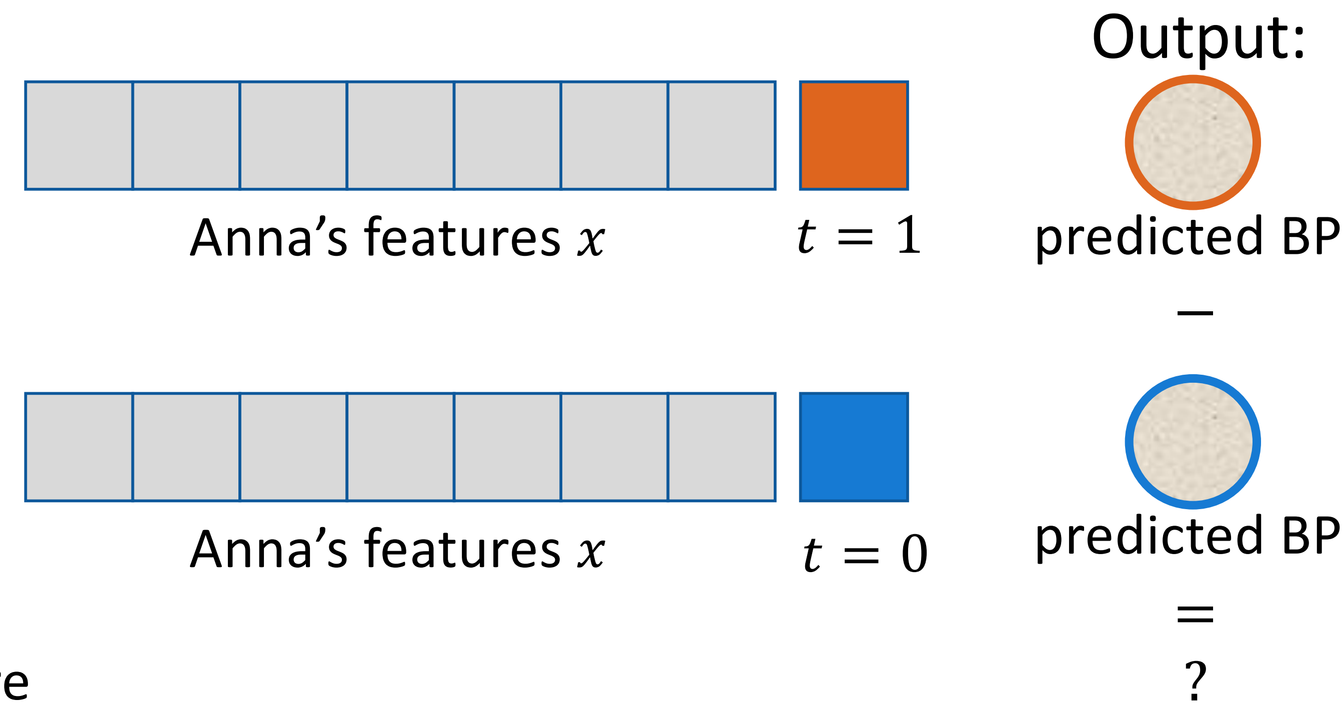
Causal inference from observational data: Medication

- Patient "Anna" comes in with hypertension
  - Asian, 54, history of diabetes, blood pressure 150/95, …
- Which of the treatments t will lower Anna's blood pressure?
  - Calcium channel blocker ($t = 1$)
  - ACE inhibitor ($t = 0$)
- Dataset of ***observational data*** medications, blood tests, past diagnoses, demographics …

Build a regression model from patient features and treatment decisions to blood pressure

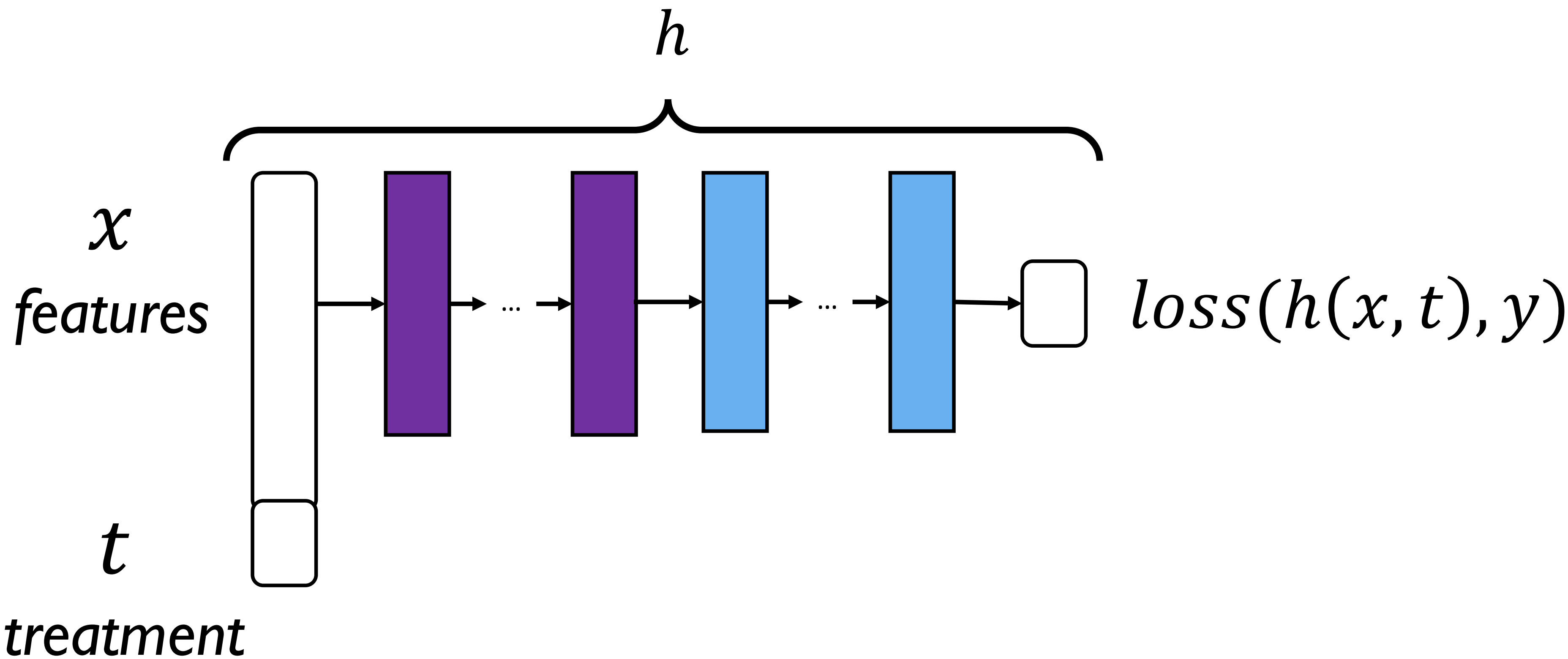
Regression modeling

- Build regression model from patient features and treatment decision to blood pressure (BP) using our observational data
- Input:

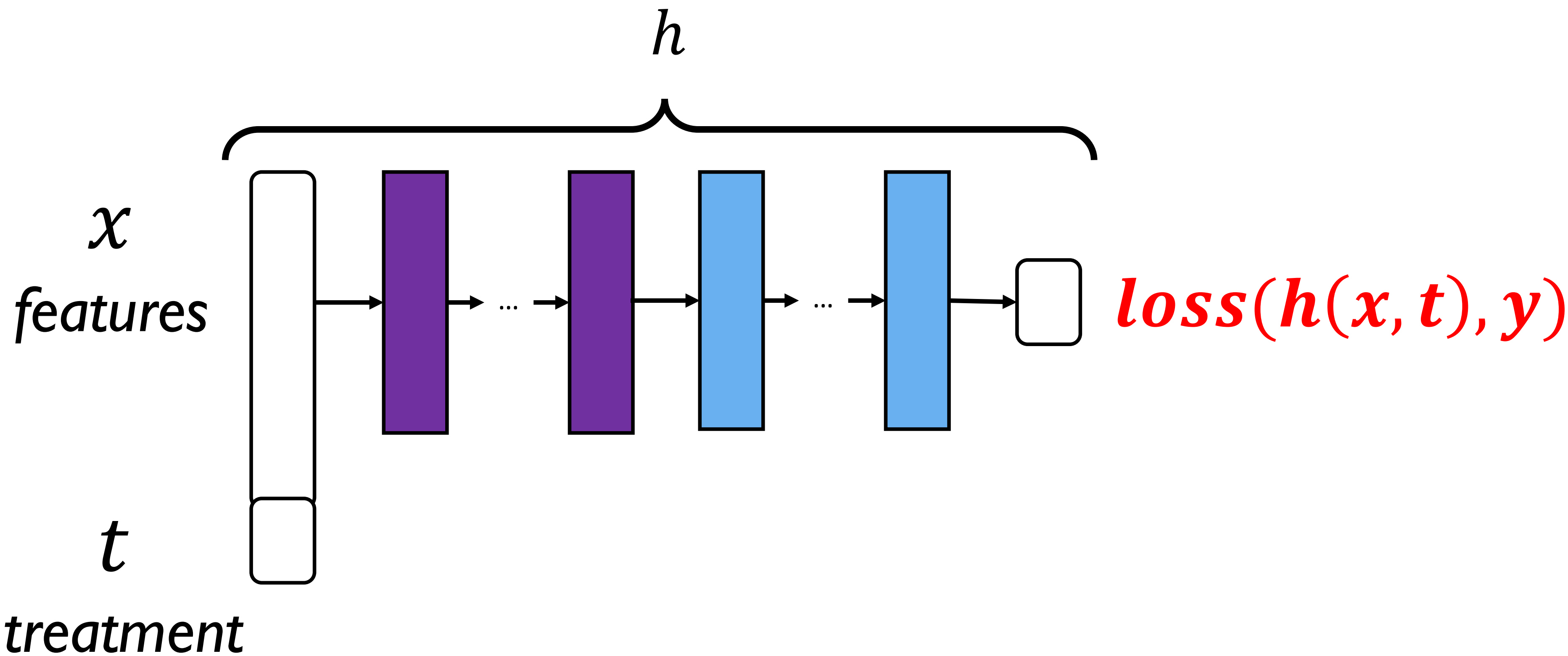


- Compare

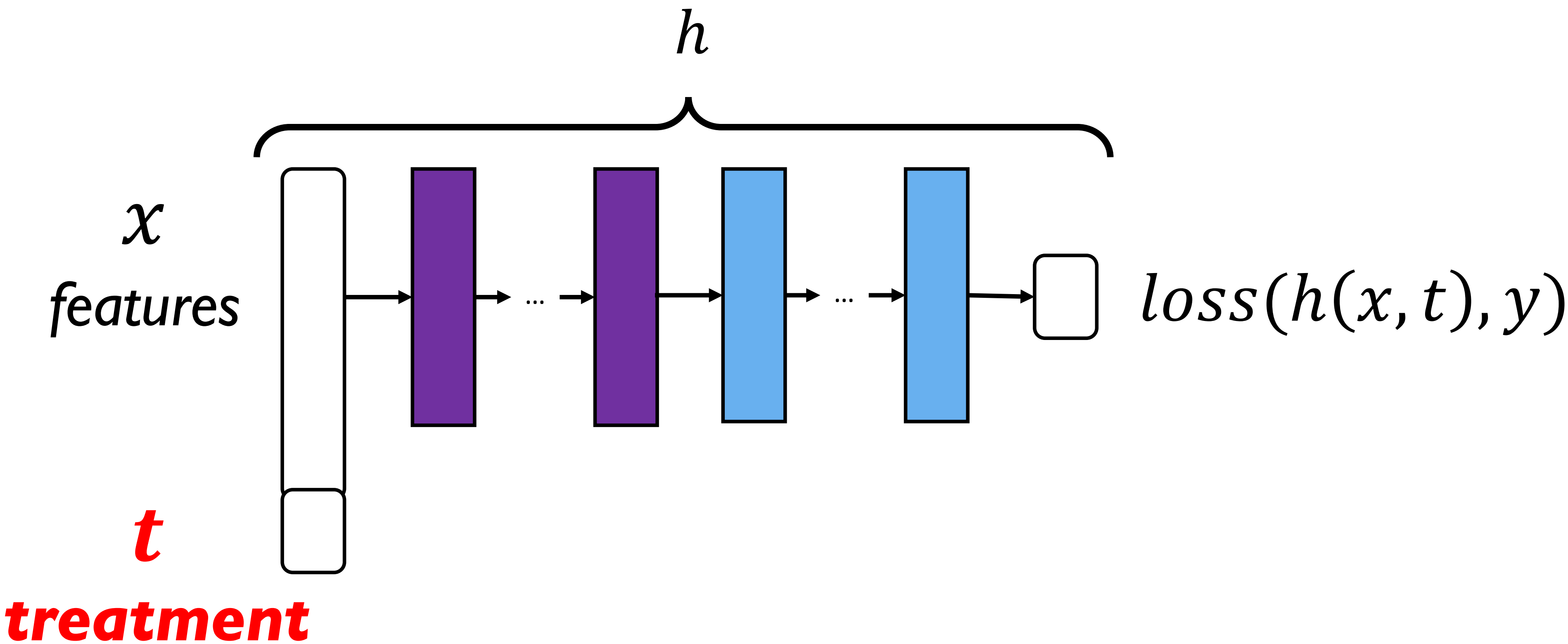
Regression modeling



Regression modeling



Regression modeling



Not supervised learning!

- This is not a classic supervised learning problem
- Supervised learning is optimized to predict outcome, not to differentiate the influence of $t = 1$ vs. $t = 0$
- What if our high-dimensional model threw away the feature of treatment t?
- Maybe there's **confounding**:
  younger patients tend to get medication $t = 1$
  older patients tend to get medication $t = 0$

Potential outcomes (Rubin & Neyman)

For every sample $x \in \mathcal{X}$, and treatment $t \in \{0,1\}$, there is a potential outcome $Y_t|x$

Blood pressure had they received treatment 1 $Y_1|x$

Blood pressure had they received treatment 0 $Y_0|x$

Individual treatment effect $\boldsymbol{ITE(x)} := \mathbb{E}[\boldsymbol{Y_1} - \boldsymbol{Y_0}|\boldsymbol{x}]$

We observe only one potential outcome, and not at random!

Example – patient blood pressure (BP)

Features: $x = (age, gender)$, $treatment$: $t \in \{0,1\}$

Factual (observed) set

(age, gender, treatment)	BP after medication
(40, F, 1)	$Y_1 = 140$
(40, M, 1)	$Y_1 = 145$
(65, F, 0)	$Y_0 = 170$
(65, M, 0)	$Y_0 = 175$
(70, F, 0)	$Y_0 = 165$

Example – patient blood pressure (BP)

Features: $x = (age, gender)$, $treatment$: $t \in \{0,1\}$

Factual (observed) set

(age, gender, treatment)	BP after medication
(40, F, 1)	$Y_1 = 140$
(40, M, 1)	$Y_1 = 145$
(65, F, 0)	$Y_0 = 170$
(65, M, 0)	$Y_0 = 175$
(70, F, 0)	$Y_0 = 165$

Counterfactual set

(age, gender, treatment)	BP after medication
(40, F, 0)	$Y_0 =?$
(40, M, 0)	$Y_0 =?$
(65, F, 1)	$Y_1 =?$
(65, M, 1)	$Y_1 =?$
(70, F, 1)	$Y_1 =?$

Example – patient blood pressure (BP)

Features: $x = (age, gender),\ treatm$

Prediction set

Factual (observed) set

(age, gender, treatment)	BP after medication
(40, F, 1)	$Y_1 = 140$
(40, M, 1)	$Y_1 = 145$
(65, F, 0)	$Y_0 = 170$
(65, M, 0)	$Y_0 = 175$
(70, F, 0)	$Y_0 = 165$

Counterfactual set

(age, gender, treatment)	BP after medication
(40, F, 0)	$Y_0 =?$
(40, M, 0)	$Y_0 =?$
(65, F, 1)	$Y_1 =?$
(65, M, 1)	$Y_1 =?$
(70, F, 1)	$Y_1 =?$

- Closely related to unsupervised domain adaptation
- No samples from the test set
- Can’t perform cross-validation!

(age, gender, treatment)	BP after medication
(40, F, 1)	$Y_1 = 140$
(40, M, 1)	$Y_1 = 145$
(65, F, 0)	$Y_0 = 170$
(65, M, 0)	$Y_0 = 175$
(70, F, 0)	$Y_0 = 165$

Prediction set

Counterfactual set

(age, gender, treatment)	BP after medication
(40, F, 0)	$Y_0 = ?$
(40, M, 0)	$Y_0 = ?$
(65, F, 1)	$Y_0 = ?$
(65, M, 1)	$Y_1 = ?$
(70, F, 1)	$Y_1 = ?$

Outline

Background

Model

Experiments

Theory

Outline

Background

Model

Experiments

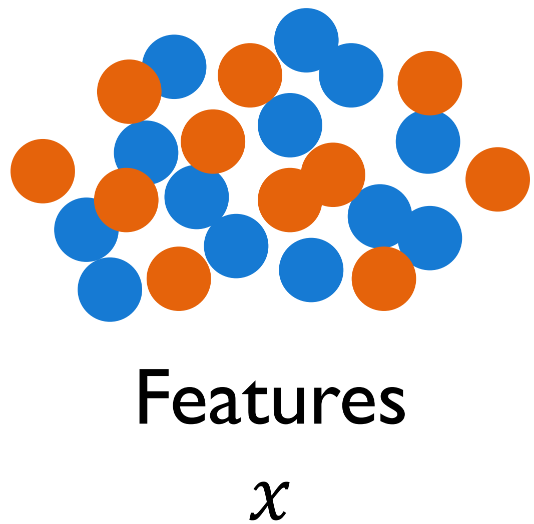
Theory

Our Work

- New neural-net based **representation learning** algorithm with explicit regularization for counterfactual estimation
- State-of-the-art on previous benchmark and on real-world causal inference task
- First error bound for estimating individual treatment effect (ITE)

When is this problem easier? Randomized Controlled Trials

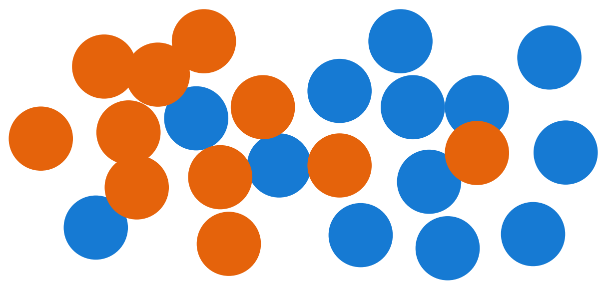
Randomized treatment → counterfactual and factual have identical distributions



- Control, $t = 0$
- Treated, $t = 1$

When is this problem harder? Observational study

Treatment assignment non-random→ counterfactual and factual have different distributions

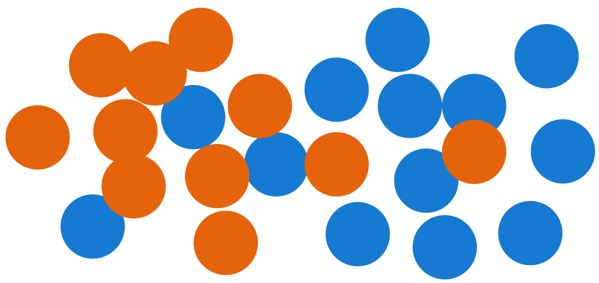


Features

x

- Control, $t = 0$
- Treated, $t = 1$

Learning more balanced representations

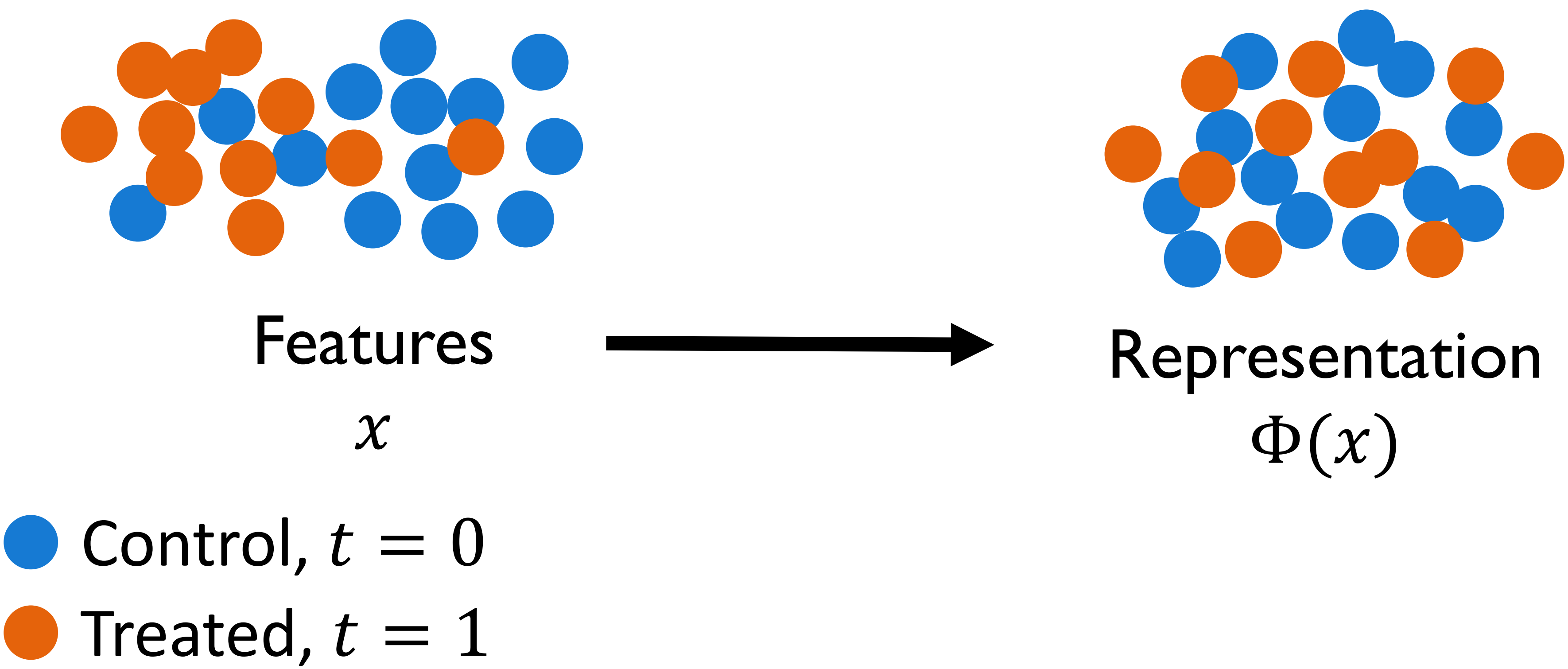


Features

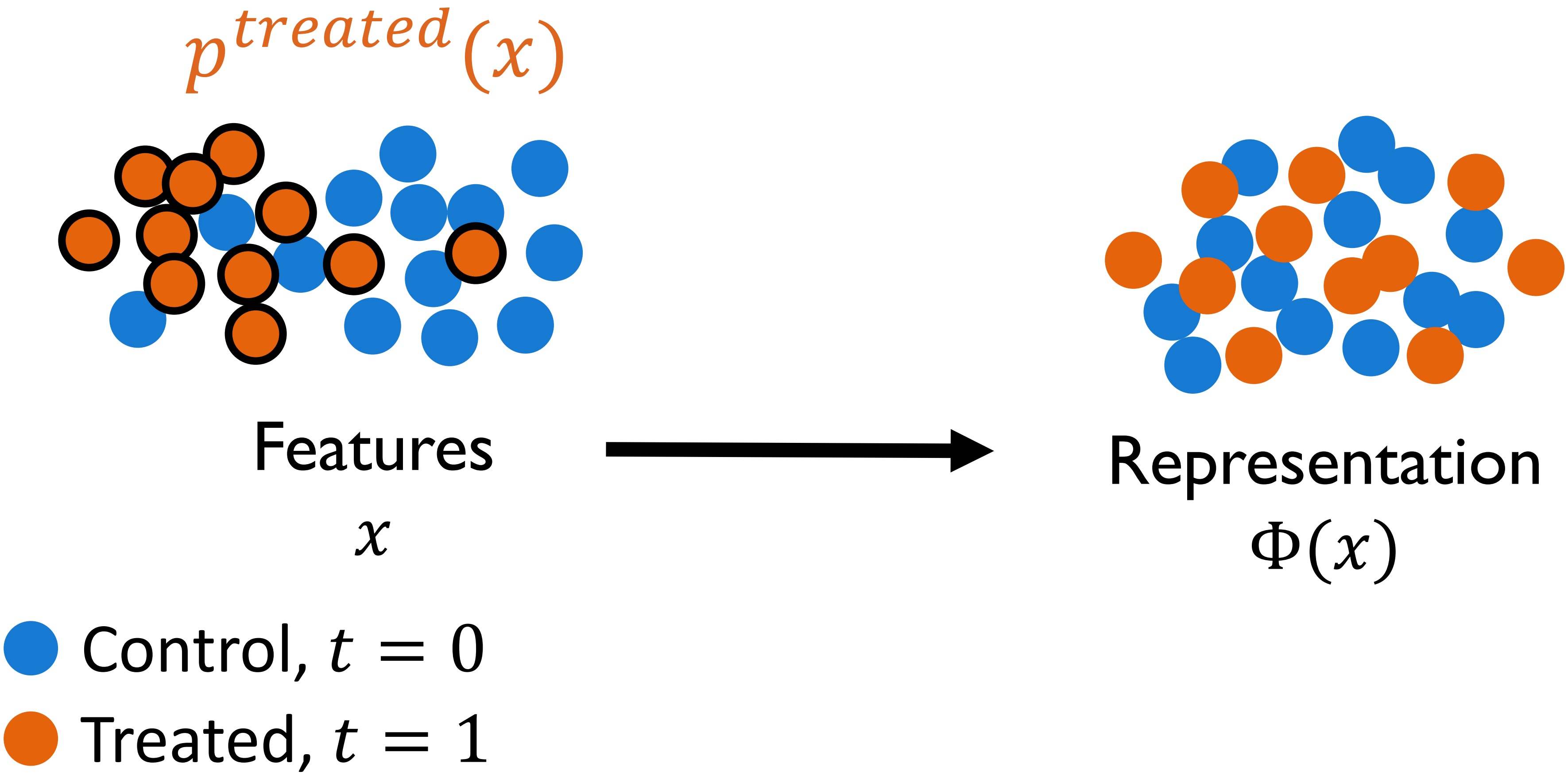
x

- Control, $t = 0$
- Treated, $t = 1$

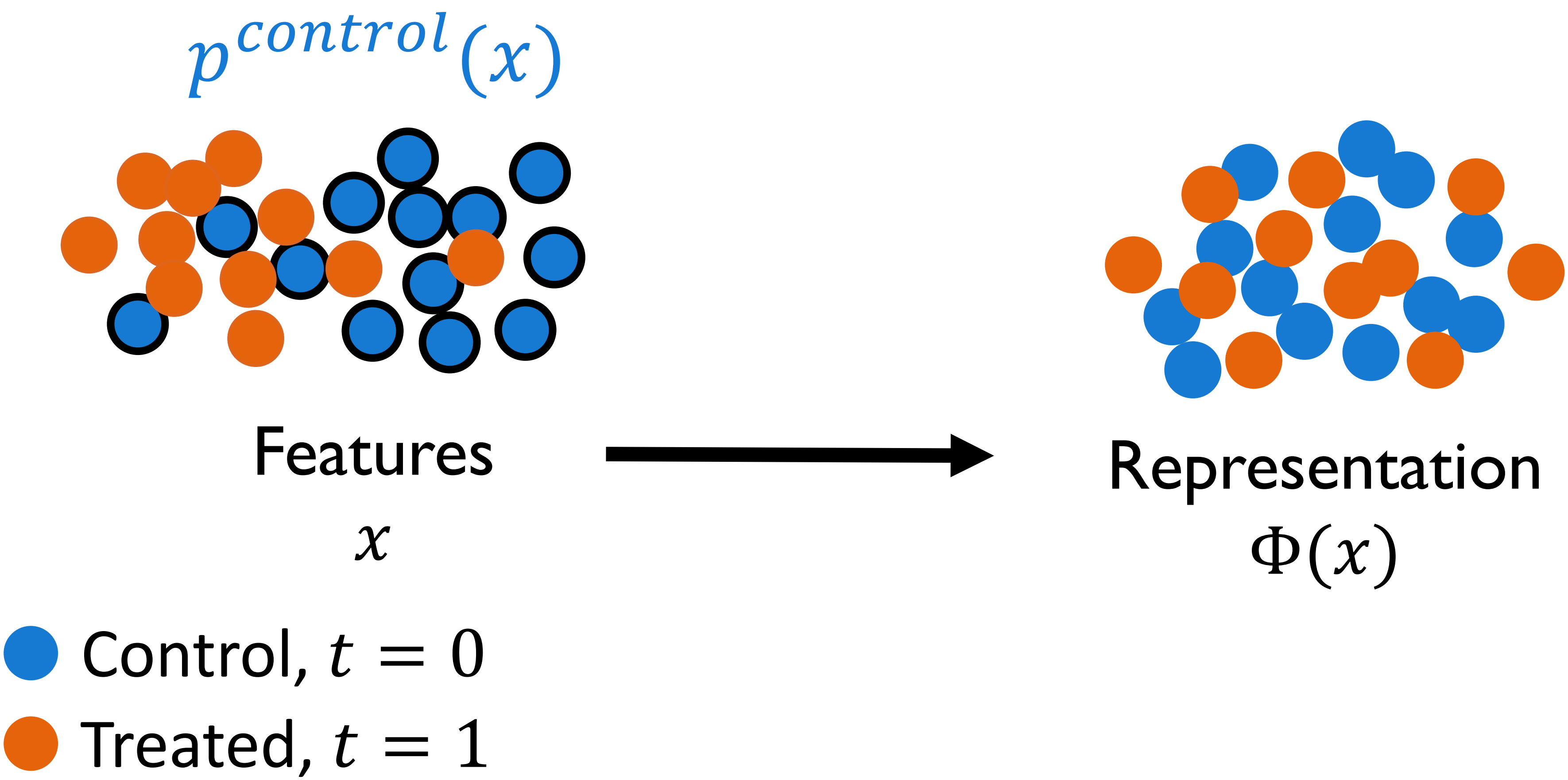
Learning more balanced representations



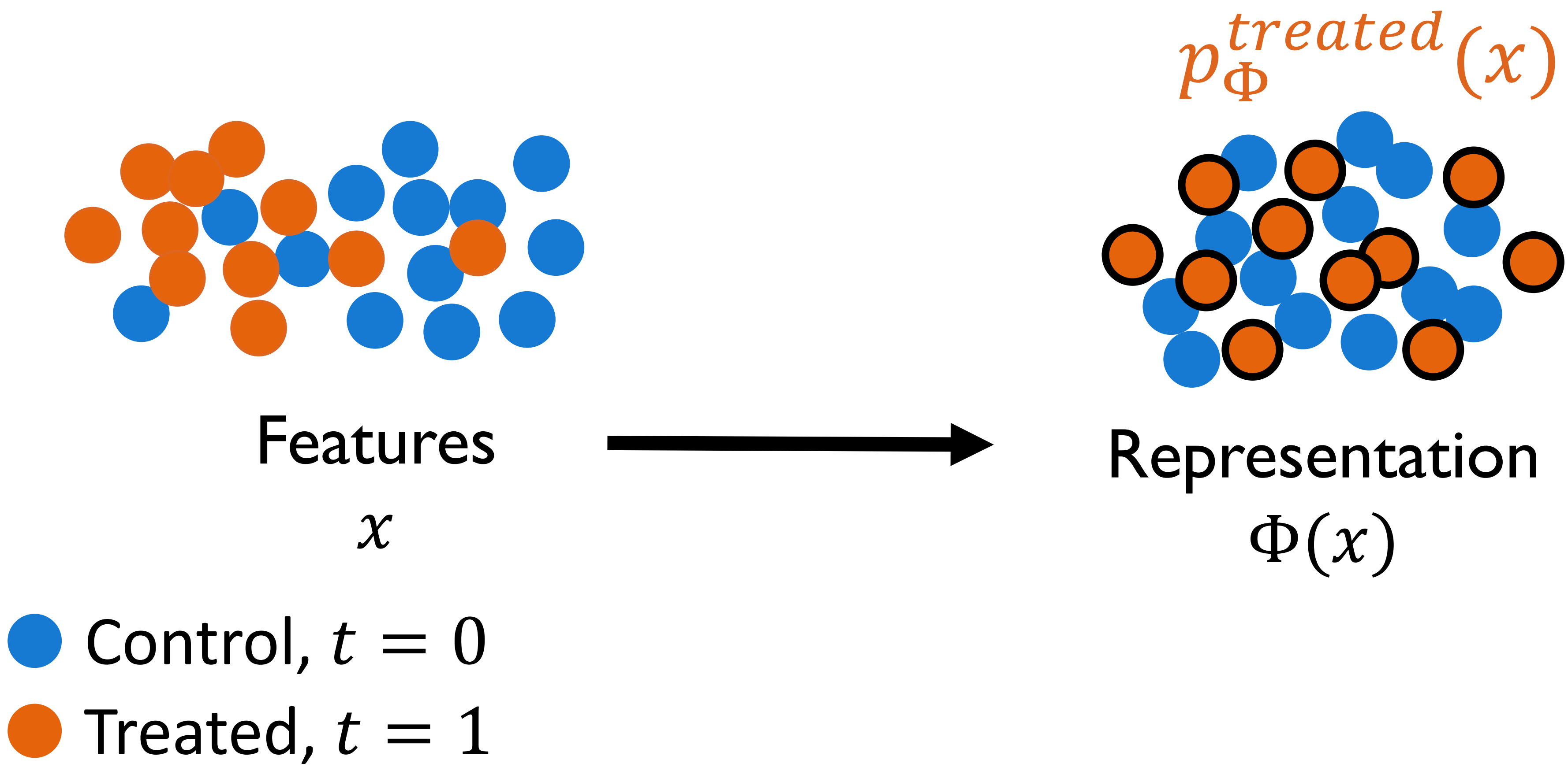
Learning more balanced representations



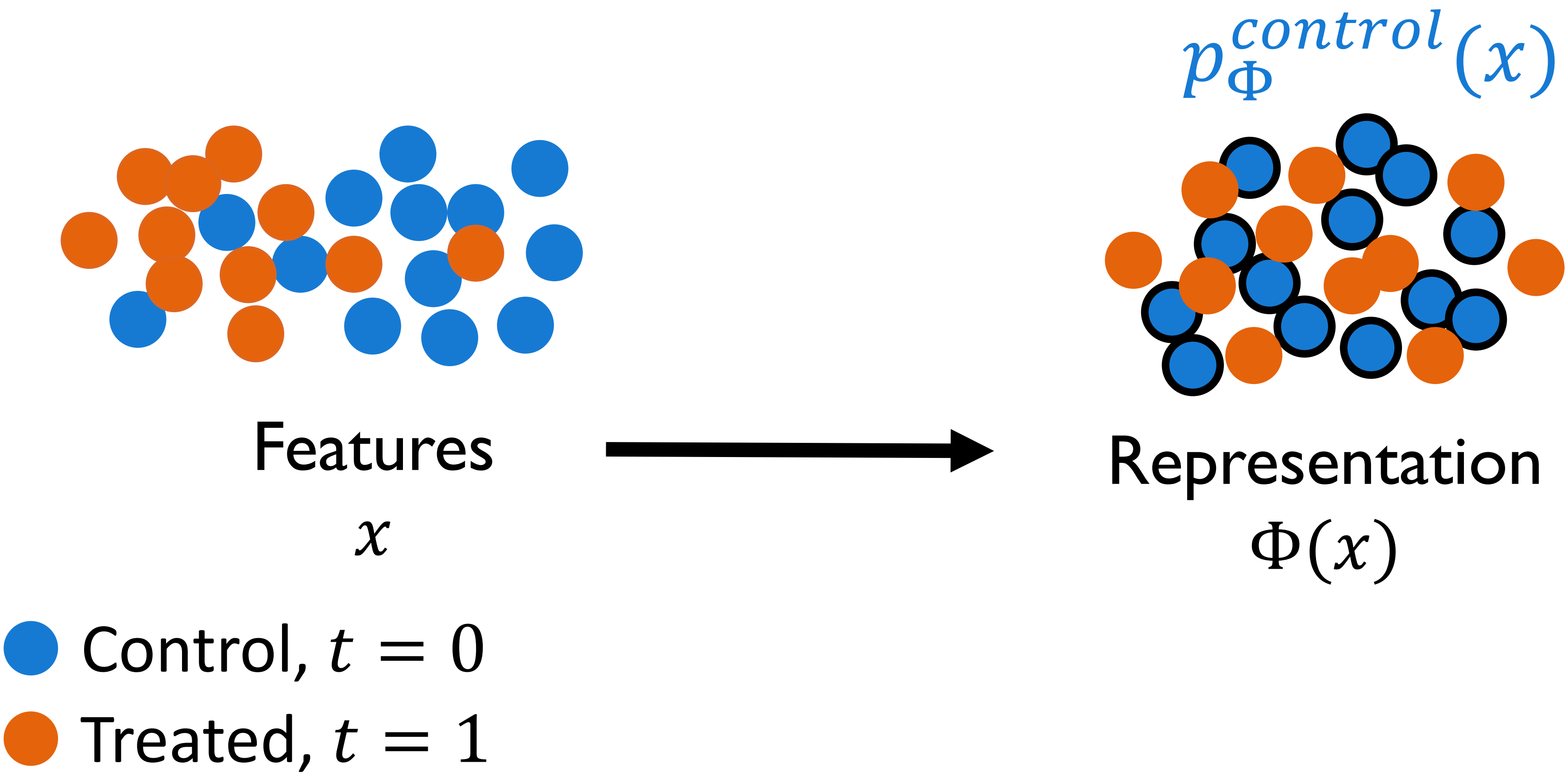
Learning more balanced representations



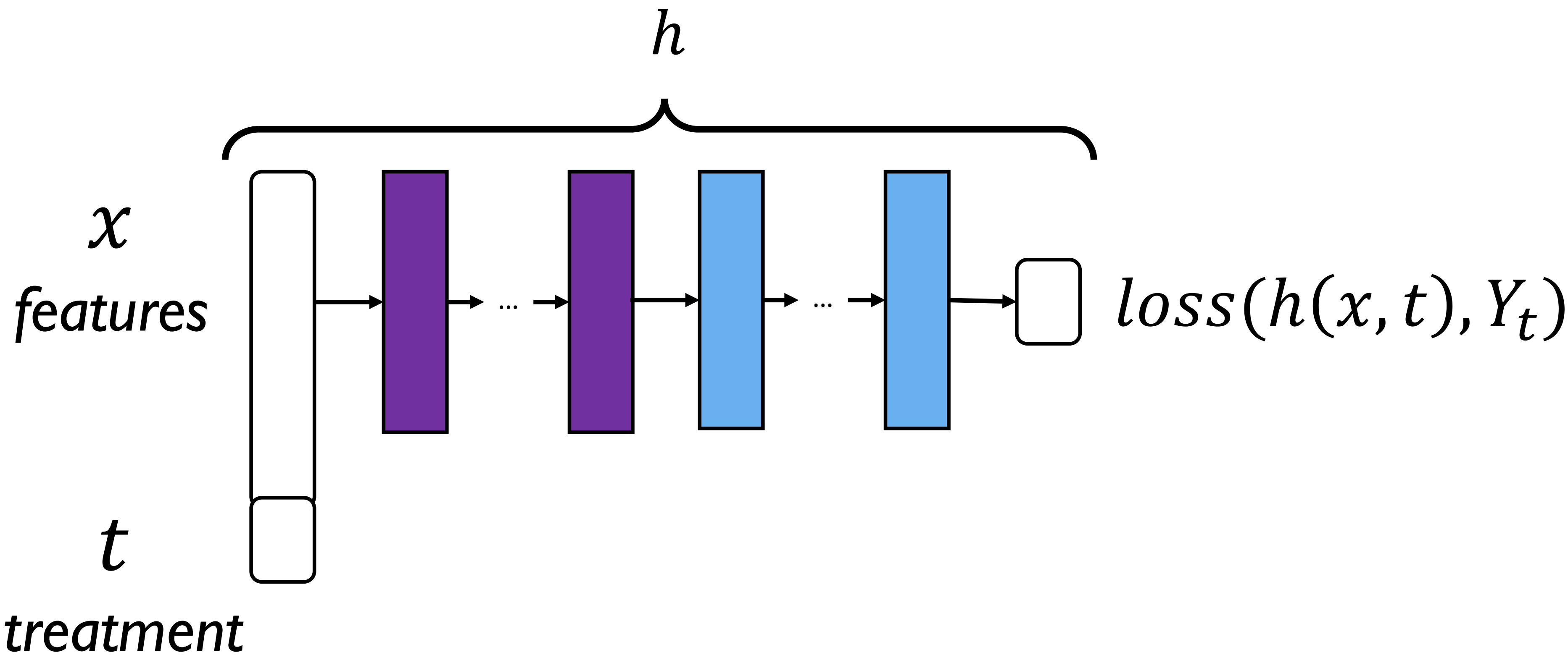
Learning more balanced representations



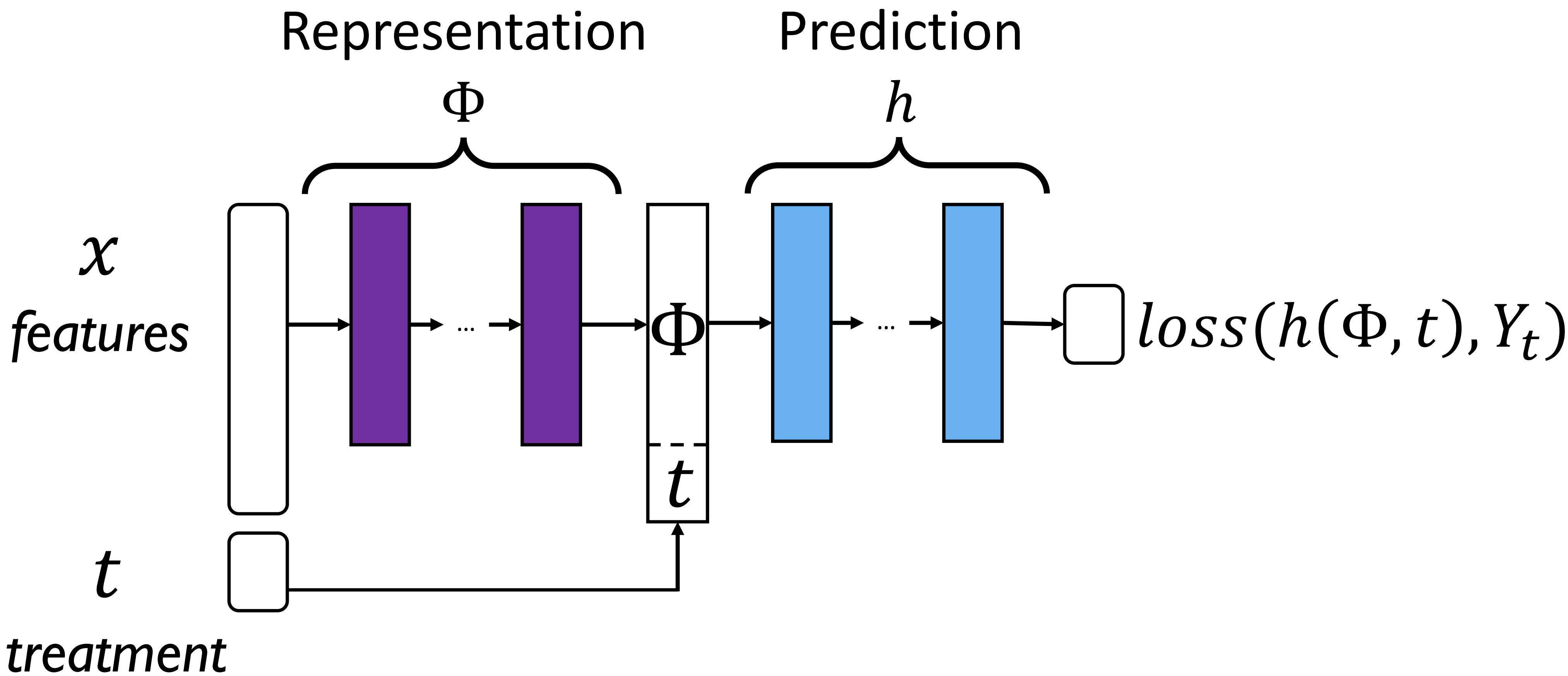
Learning more balanced representations



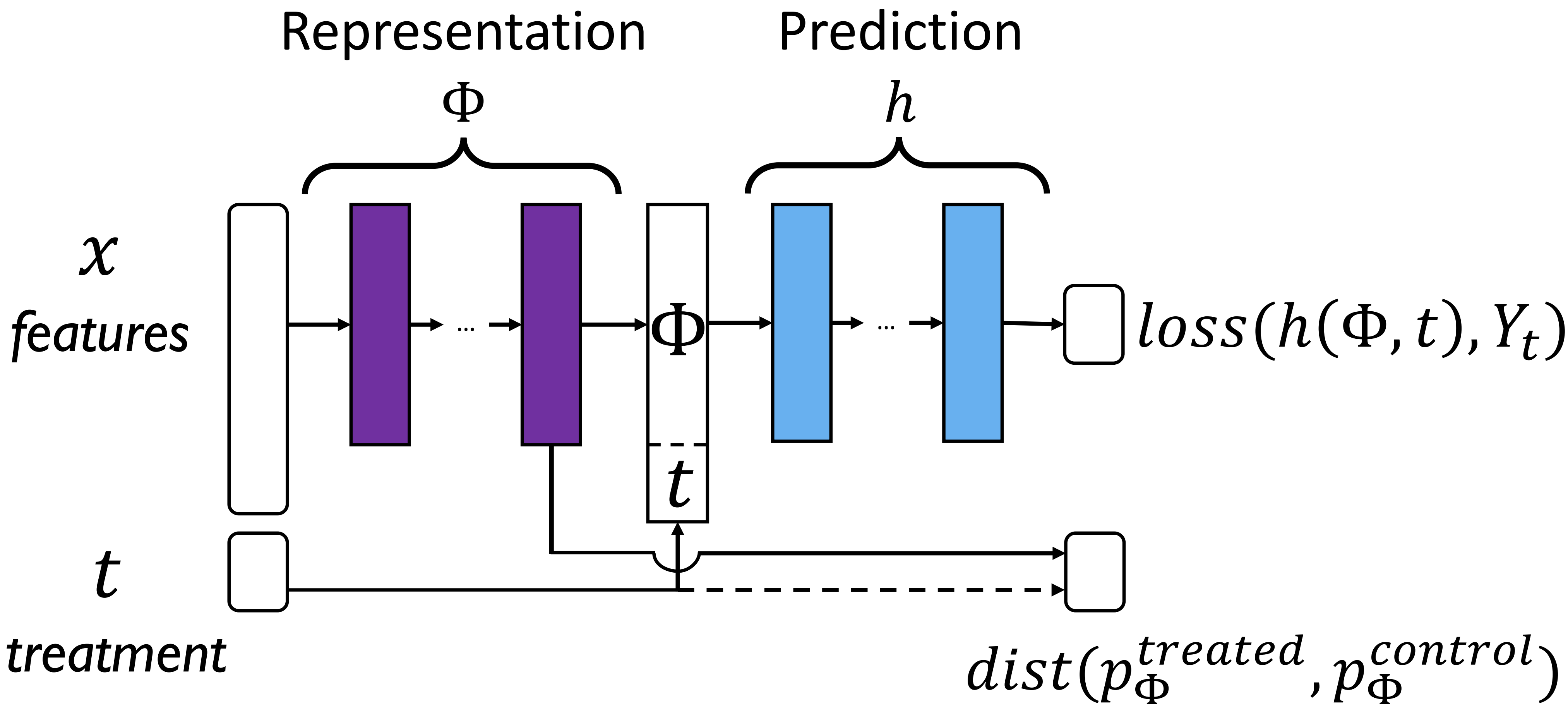
Naïve Neural Network for estimating individual treatment effect (ITE)



Vanilla Neural Network for Counterfactual Regression (CFR)



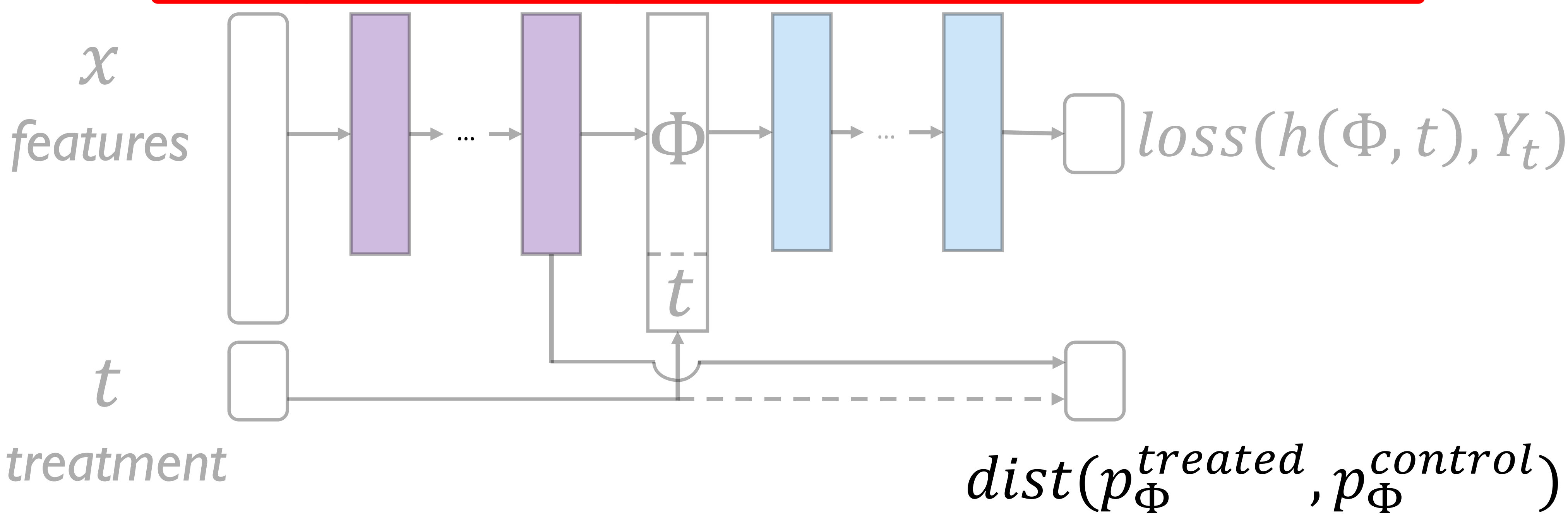
Balancing Neural Network for Counterfactual Regression (CFR)



$dist(p_{\Phi}^{treated}, p_{\Phi}^{control})$:

MMD distance (Gretton et al. 2012)

Wasserstein distance (Villani 2008, Cuturi 2013)



Bala sion

$dist(p_\Phi^{treated}, p_\Phi^{control})$:

MMD distance (Gretton et al. 2012)
Wasserstein distance (Villani 2008, Cuturi 2013)

Inspired by Domain Adversarial Networks (Ganin et al., 2016):

`(source domain, target domain)` →
`(treated population, control population)`

t

treatment

$dist(p_\Phi^{treated}, p_\Phi^{control})$

Outline

Background

Model

Experiments

Theory

Outline

Background

Model

Experiments

Theory

Evaluating counterfactual inference

Train-test paradigm breaks
No observations from the counterfactual “test” set

Can’t do cross-validation for hyper-parameter selection

1) Simulated data: IHDP (Hill, 2011)

2) Real data: National Supported Work study (LaLonde, 1986, Todd & Smith 2005)
The effect of job training on employment and income

Observational study with a *randomized controlled trial subset*

Evaluating counterfactual inference

Train-test paradigm breaks
No observations from the counterfactual "test" set

Can't do cross-validation for hyper-parameter selection

1) Simulated data: IHDP (Hill, 2011)

2) Real data: National Supported Work study (LaLonde, 1986, Todd & Smith 2005)
The effect of job training on <u>employment</u> and income

Observational study with a *randomized controlled trial subset*
3212 samples, 8 features incl. education and previous income

Evaluating models with randomized controlled trials data

- We can't directly evaluate individual treatment effect (ITE) error because we never see the counterfactual

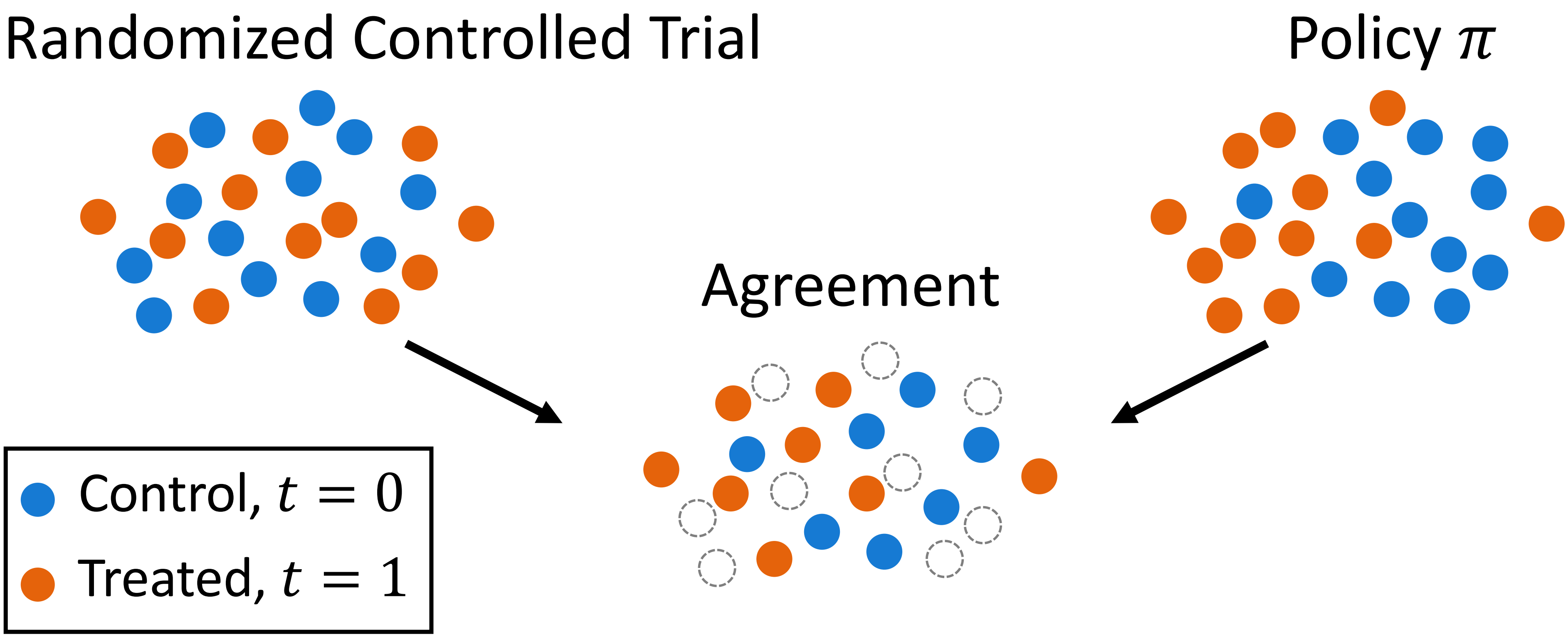
- Every ITE estimator implies a policy
  $\widehat{ITE}(x) = f(x)$

  Policy $\pi_{f,\lambda}: \mathcal{X} \rightarrow \{0,1\}$
  Treat all persons x with $f(x) > \lambda$, for threshold λ

- Every policy π has a policy-value:

$$\mathbb{E}[Y_1 | \pi(x) = 1] p(\pi = 1) + \mathbb{E}[Y_0 | \pi(x) = 0] p(\pi = 0)$$

Evaluating model performance using randomized data (off-policy evaluation)



Policy value: $\mathbb{E}[Y_1|\pi(x) = 1]p(\pi = 1) + \mathbb{E}[Y_0|\pi(x) = 0]p(\pi = 0)$

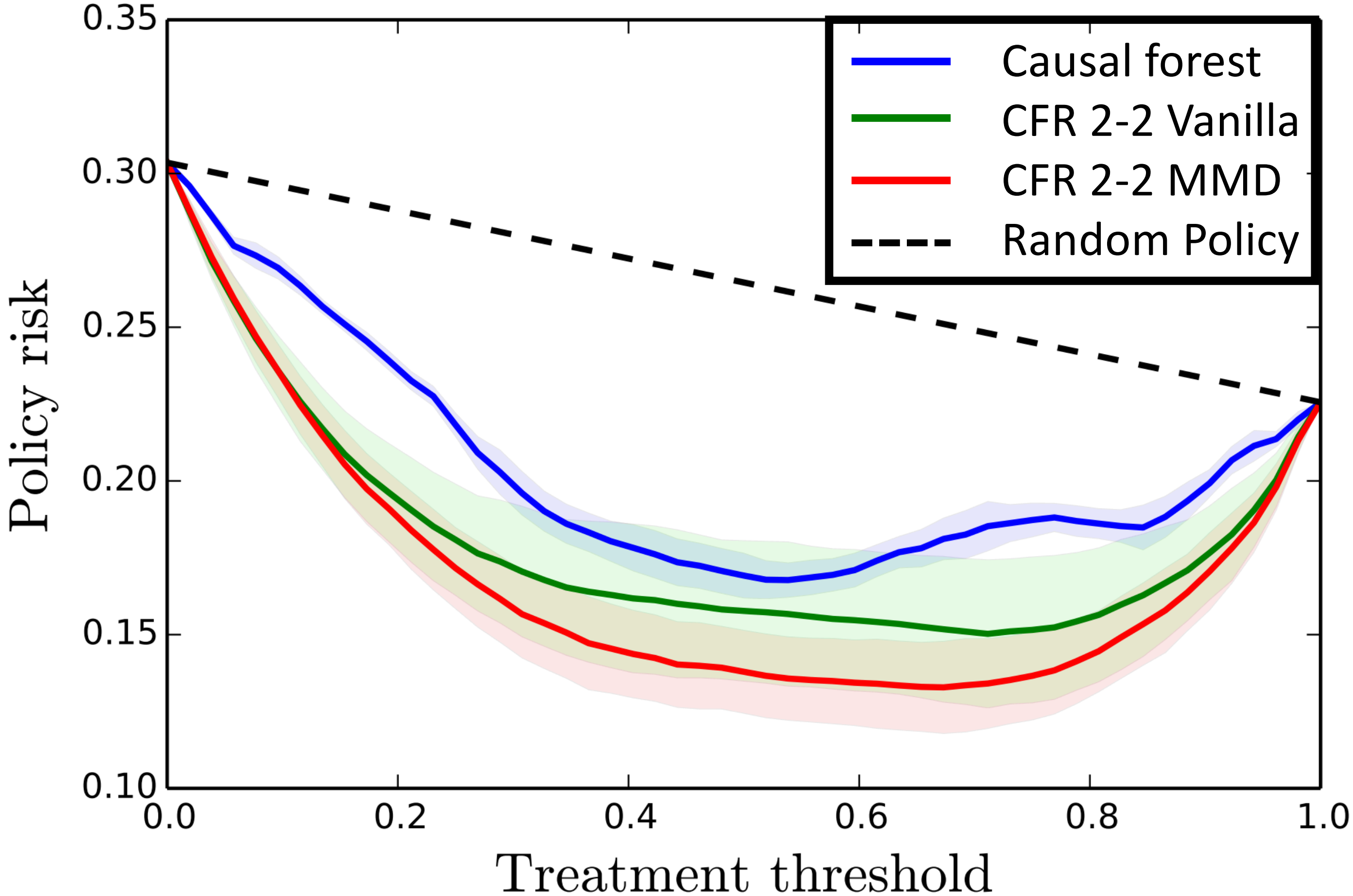
Experimental results – National Supported Work Study

- National Supported Work: randomized trial embedded in an observational study
- *Policy risk estimated on randomized subsample*
- CFR-2-2: our model, with 2 layers before Φ and 2 layers after Φ

Method	**Policy risk (std)**
ℓ_1-reg. logistic regression	0.23±0.00
BART (Chipman, George & McCulloch, 2010)	0.24±0.01
Causal forests (Wager & Athey, 2015)	0.17±0.006
CFR-2-2 Vanilla	0.16±0.02
CFR-2-2 Wasserstein	0.15±0.02
CFR-2-2 MMD	**0.13**±0.02

Lower is better

Experimental results – National Supported Work Study



Lower is better

Outline

Background

Model

Experiments

Theory

Outline

Background

Model

Experiments

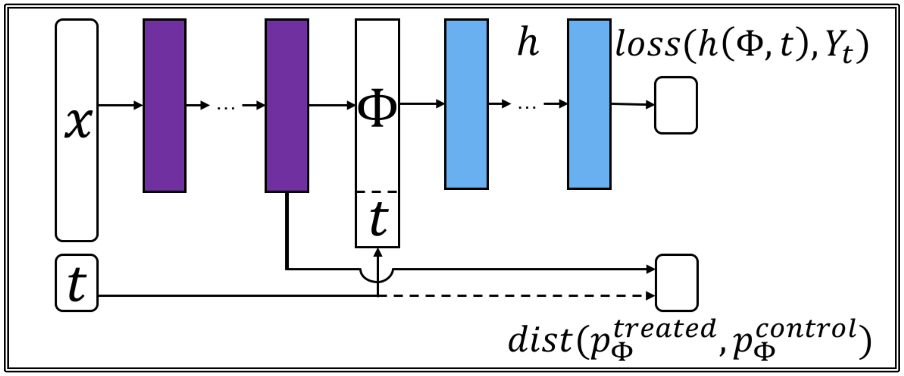
Theory

Theory of causal effect inference

- Standard results in statistics: asymptotic rate of convergence to true average effect
  - Assumptions: we know true model (consistency)

- Our result: generalization error bound for individual-level inference
  - Assumptions: true model lies within large model family, e.g. bounded Lipschitz functions

Theorem (informal)



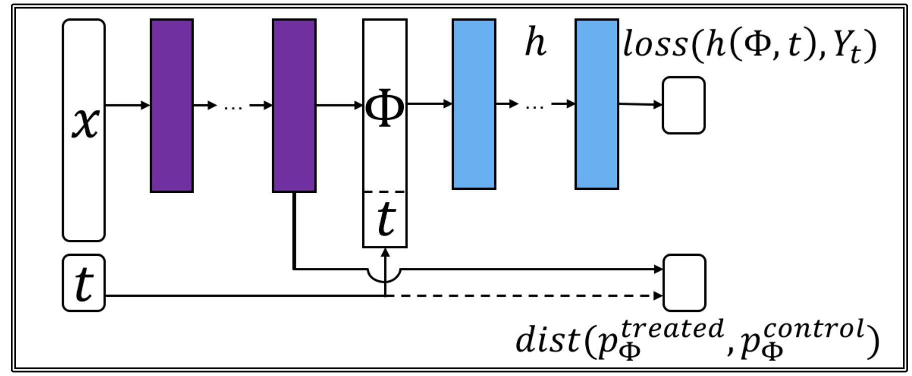
- Let $\hat{Y}_t^{\Phi,h}(x) = h(\Phi(x), t)$ for $t = 0,1$
- $\widehat{ITE}^{\Phi,h}(x) := \hat{Y}_1^{\Phi,h}(x) - \hat{Y}_0^{\Phi,h}(x)$

- If "strong ignorability" holds, and if $dist$ is "nice" with respect to the true potential outcomes Y_0 and Y_1 and the representation Φ, then for all normalized Φ and h:

$$\mathbb{E}_x\left[error\left(\widehat{ITE}^{\Phi,h}(x)\right)\right] \leq$$
$$2 \cdot \mathbb{E}_{x,t}\left[error\left(\hat{Y}_t^{\Phi,h}(x)\right)\right] + dist(p_\Phi^{treated}, p_\Phi^{control})$$

Theorem (informal)

- Let $\hat{Y}_t^{\Phi,\mathrm{h}}(x) = h(\Phi(x), t)$ for $t = 0,1$
- $\widehat{ITE}^{\Phi,\mathrm{h}}(x) := \hat{Y}_1^{\Phi,\mathrm{h}}(x) - \hat{Y}_0^{\Phi,\mathrm{h}}(x)$

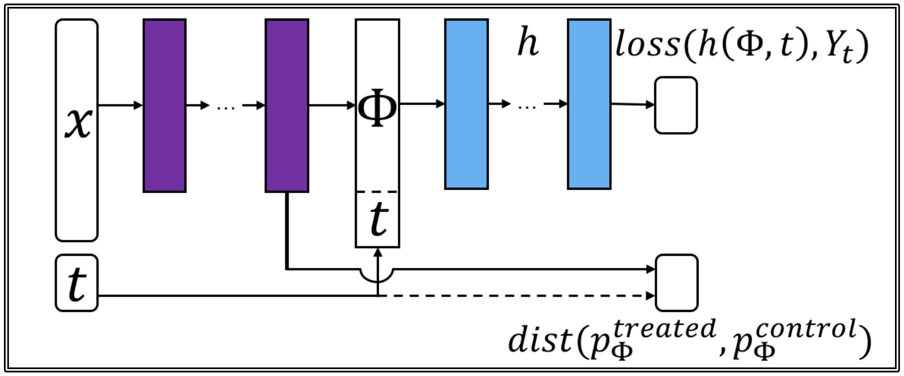


- If “strong […]s, and if $dist$ is “nice” with respect to the true p[…]s Y_0 and Y_1 and the representation Φ, then for a[…]nd h:

Expected error in estimating ITE

$$\mathbb{E}_x\left[error\left(\widehat{ITE}^{\Phi,h}(x)\right)\right] \leq$$
$$2 \cdot \mathbb{E}_{x,t}\left[error\left(\hat{Y}_t^{\Phi,h}(x)\right)\right] + dist(p_\Phi^{treated}, p_\Phi^{control})$$

Theorem (informal)



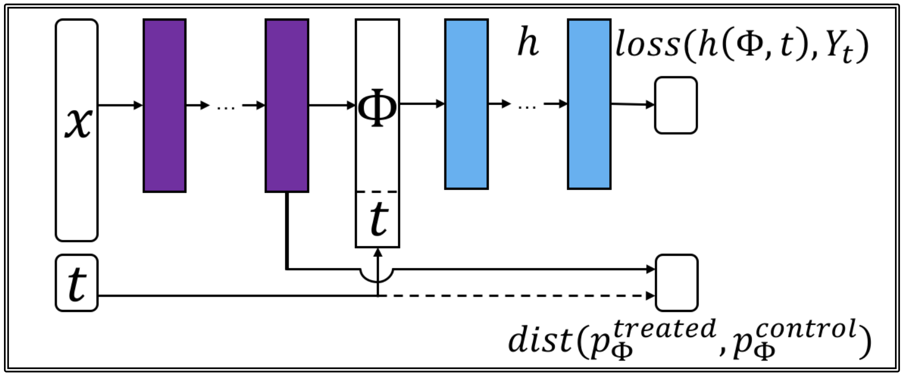
- Let $\hat{Y}_t^{\Phi,\mathrm{h}}(x) = h(\Phi(x),t)$ for $t = 0{,}1$
- $\widehat{ITE}^{\Phi,\mathrm{h}}(x) := \hat{Y}_1^{\Phi,\mathrm{h}}(x) - \hat{Y}_0^{\Phi,\mathrm{h}}(x)$

- If "strong ignorability" holds, and if $dist$ is "nice" with respect to the true potential outcomes Y_0 and Y_1 and the representation Φ, then for all normalized Φ and h:

$$\mathbb{E}_x\left[error\left(\widehat{ITE}^{\Phi,h}(x)\right)\right] \leq$$
$$2 \cdot \underbrace{\mathbb{E}_{x,t}\left[error\left(\hat{Y}_t^{\Phi,h}(x)\right)\right]}_{\text{"supervised learning generalization error"}} + dist(p_\Phi^{treated}, p_\Phi^{control})$$

"supervised learning generalization error"

Theorem (informal)



- Let $\hat{Y}_t^{\Phi,h}(x) = h(\Phi(x), t)$ for $t = 0,1$
- $\widehat{ITE}^{\Phi,h}(x) := \hat{Y}_1^{\Phi,h}(x) - \hat{Y}_0^{\Phi,h}(x)$

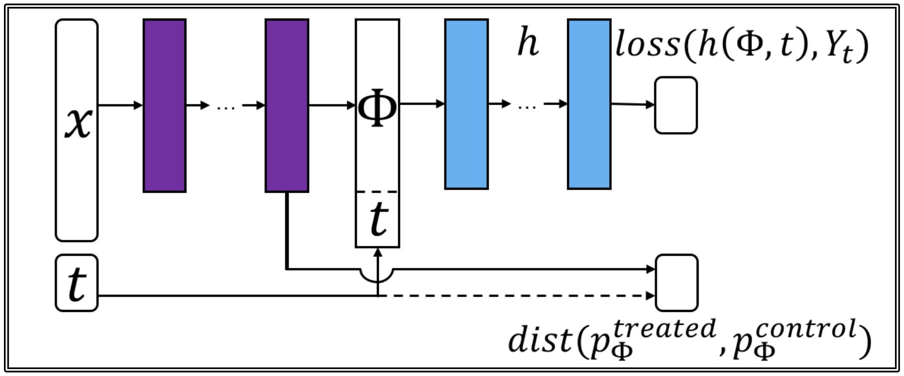
- If "strong ignorability" holds, and if $dist$ is "nice" with respect to the true potential outcomes Y_0 and Y_1 and the representation Φ, then for all normalized Φ and h:

$$\mathbb{E}_x\left[error\left(\widehat{ITE}^{\Phi,h}(x)\right)\right] \leq$$
$$2 \cdot \mathbb{E}_{x,t}\left[error\left(\hat{Y}_t^{\Phi,h}(x)\right)\right] + \underbrace{dist(p_\Phi^{treated}, p_\Phi^{control})}_{\text{Distance between } \Phi\text{-induced distributions}}$$

Distance between Φ-induced distributions

Theorem (informal)

- Let $\hat{Y}_t^{\Phi,\mathrm{h}}(x) = h(\Phi(x), t)$ for $t = 0,1$
- $\widehat{ITE}^{\Phi,\mathrm{h}}(x) := \hat{Y}_1^{\Phi,\mathrm{h}}(x) - \hat{Y}_0^{\Phi,\mathrm{h}}(x)$



We minimize upper bound with respect to Φ and h

$$\mathbb{E}_x\left[error\left(\widehat{ITE}^{\Phi,h}(x)\right)\right] \leq$$
$$2 \cdot \mathbb{E}_{x,t}\left[error\left(\hat{Y}_t^{\Phi,h}(x)\right)\right] + dist(p_\Phi^{treated}, p_\Phi^{control})$$

Summary

- Estimating Individual Treatment Effect is different from supervised learning
  - Bears strong connections to domain adaptation
- We give new representation learning algorithms for estimating Individual Treatment Effect
  - Use the MMD and Wasserstein distributional distances
- Experiments show our method is competitive or better than state-of-the-art
- We give a new error bound for estimating Individual Treatment Effect

Acknowledgments: Justin Chiu (FAIR), Marco Cuturi (ENSAE / CREST), Jennifer Hill (NYU), Aahlad Manas (NYU), Sanjong Misra (U. Chicago), Esteban Tabak (NYU) and Stefan Wager (Columbia)

Thank you!

- Fredrik D. Johansson, Uri Shalit, David Sontag
  "*Learning Representations for Counterfactual Inference*"
  ICML 2016
- Uri Shalit, Fredrik D. Johansson, David Sontag
  "*Estimating individual treatment effect: generalization bounds and algorithms*"
  arXiv:1606.03976