Learning Representations for Counterfactual Inference

Fredrik Johansson[1], Uri Shalit[2], David Sontag[2]

[1] CHALMERS UNIVERSITY OF TECHNOLOGY [2] NYU

Counterfactual inference

Patient “Anna” comes in with *hypertension.* She is 50 years old, Asian and has a history of diabetes. What would Anna’s blood pressure (BP) be after receiving medication A?

After receiving medication A, Anna’s BP is somewhat better.

What would have happened if she instead received medication B?

Counterfactual!

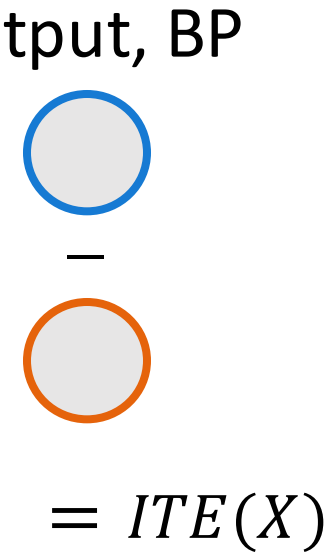
We have a dataset of 1 000 000 patients with hypertension that received medication A or B. *Not at random!*

Machine learning for counterfactuals

Build regression model from patient features to blood pressure (BP)

Focus on difference in outcomes, the individual treatment effect (ITE)!

Not so much the predicted outcome



Main contributions

- Reduce counterfactual inference to domain adaptation
- Give representation learning algorithm for minimizing counterfactual error
- Bound difference between counterfactual and factual error
- Show neural nets can be used for counterfactual inference

Potential outcomes [Rubin'74]

For every sample $x \in \mathcal{X}$, and treatment $t \in \mathcal{T}$, there is a potential outcome $Y_t(x)$. We observe *only one!*

Blood pressure after medication A $Y_0(x)$ Control

Blood pressure after medication B $Y_1(x)$ Treated

Individual treatment effect: $ITE(x) = Y_1(x) - Y_0(x)$

Example – patient blood pressure (BP)

Features: x = (age, gender), Treatment: t = (medication A or B)

Factual (observed) set

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

Counterfactual set

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

Causal inference as domain adaptation

Factual = Source domain

$p_F(x,t) = p_F(x)p_F(t|x)$

Labeled with y_i

Control

Treated

$p_{CF}(x,t) \coloneqq p_F(x)p_F(1-t|x)$

Unlabeled

Counterfactual = Target domain

Our key insight

- Build on connection to domain adaptation*
- *Learn representations* that

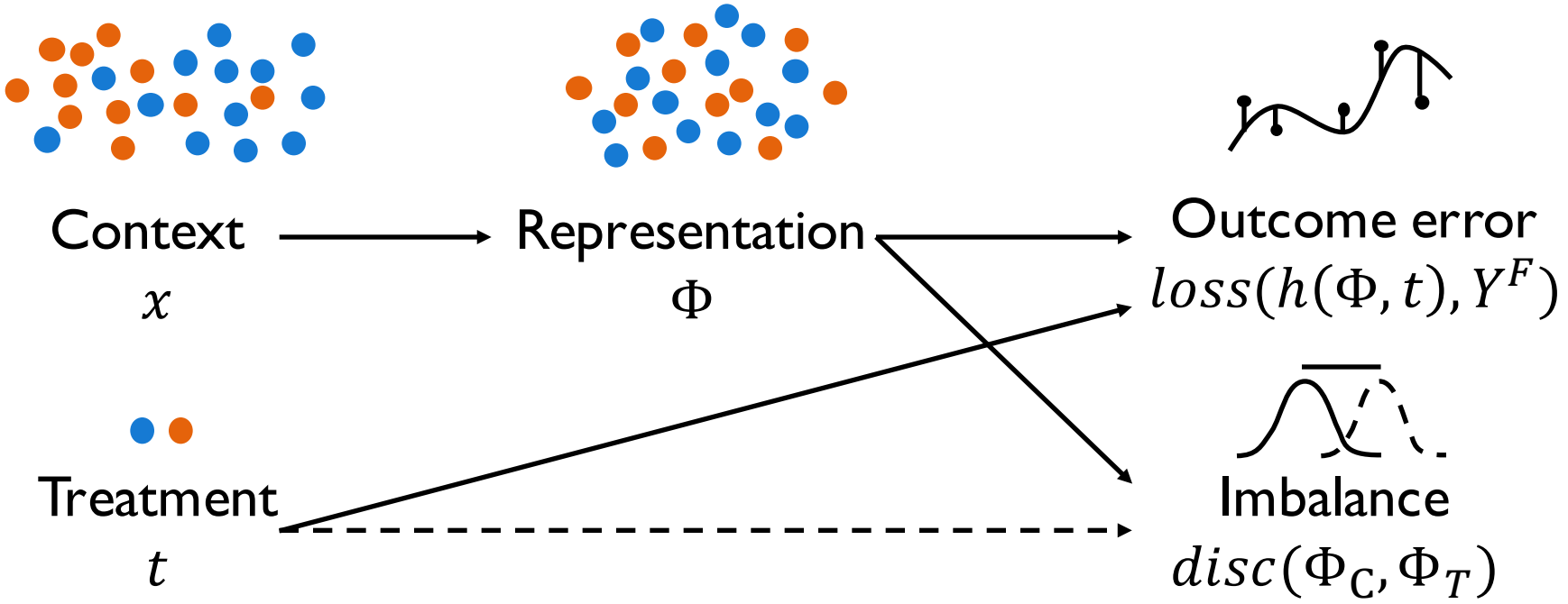
a) Enable good factual prediction

b) Limit gap between factual and counterfactual

Make treated and control patients look more similar!

*See e.g. Ganin et al., 2015; Zemel et al., 2013

Learning balanced representations



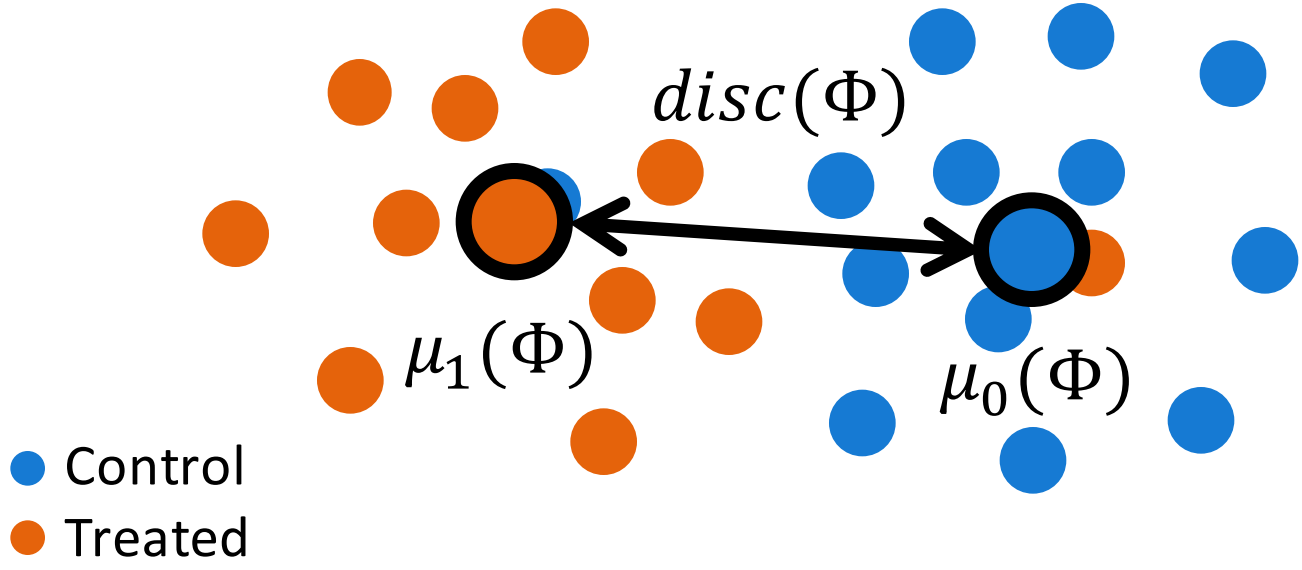
Measuring imbalance

Discrepancy: distance between distributions w. r. t. loss function L and hypothesis space $\mathcal{H}$ (Mansour et al., 2009)

When $\mathcal{H}$ is the set of *linear* hypotheses

$$disc_{\mathcal{H}}(\mathrm{P}, \mathrm{Q}) = \|\mu(P) - \mu(Q)\|_2$$

Measuring imbalance

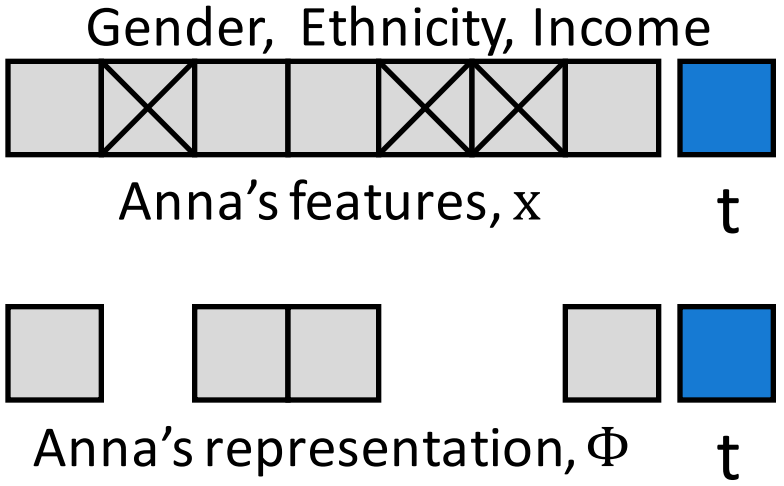


Balancing Linear Regression (BLR)

Select & reweight covariates for linear regression that are either

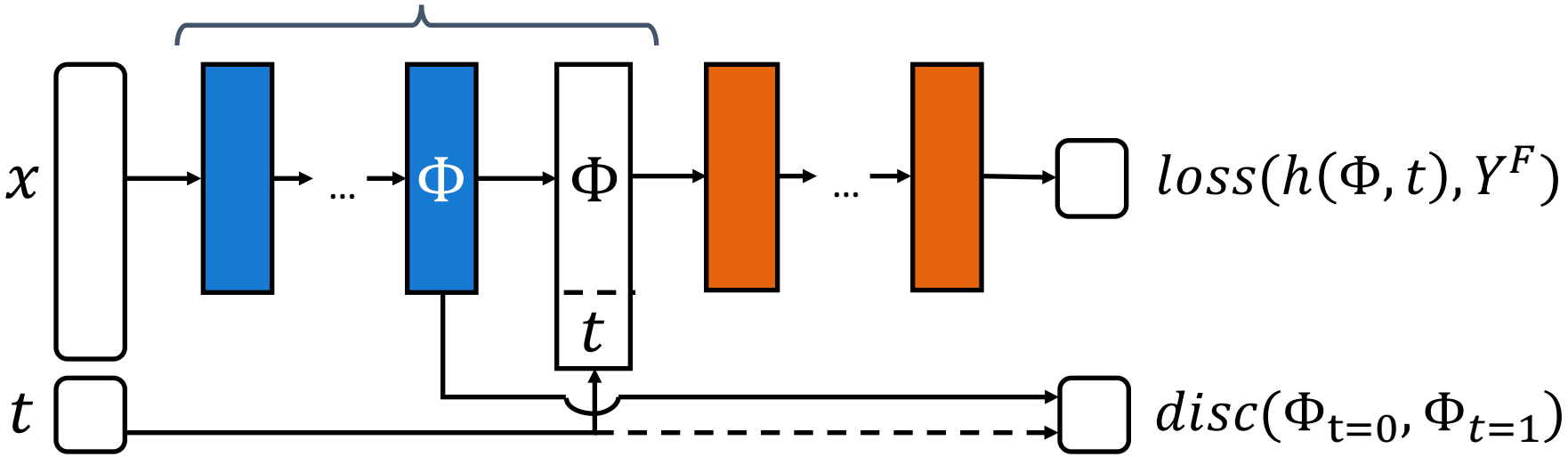
a) Predictive

b) Balanced



Balancing Neural Network (BNN)

Learned to be *balanced* for treated and control



Algorithm. Learning representation Φ

1. Learn a representation Φ by minimizing factual loss with discrepancy between treated and control as regularizer
2. Fit ridge regression hypothesis h to the learned Φ

Theorem. Bounding counterfactual error

We bound the difference between factual and counterfactual error. Building on (Cortes & Mohri, 2014)

Expected ℓ_1-loss on the factual

Distance to counterfactual neighbors

$$(\epsilon_{CF} - \epsilon_F)^2 \leq a \cdot loss_F(\Phi, h) + b \cdot disc(\Phi) + c \cdot \eta_{NN}$$

Expected squared error on factual and counterfactual

Imbalance between control and treated

Evaluating counterfactual inference

"Train-test paradigm breaks" (S. Athey)
No observations from the test set

Can't do cross-validation for hyperparameter selection

Evaluate on simulated data: IHDP (Hill, 2011)

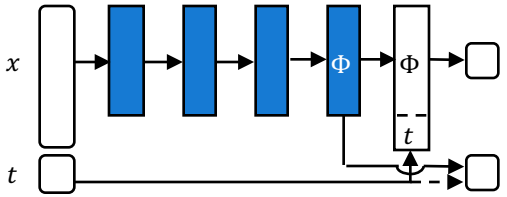
Infant Health & Development Program

Real features and simulated log-linear outcome

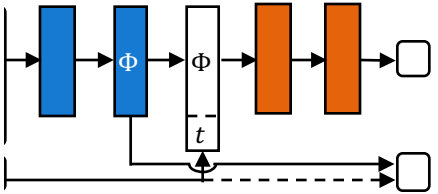
Results on IHDP

	Error in individual effect estimate	Error in average effect estimate
LINEAR OUTCOME		
OLS	4.6 ± 0.2	0.7 ± 0.0
DOUBLY ROBUST	3.0 ± 0.1	0.2 ± 0.0
LASSO + RIDGE	2.8 ± 0.1	0.2 ± 0.0
BLR	2.8 ± 0.1	0.2 ± 0.0
BNN-4-0	3.0 ± 0.0	0.3 ± 0.0

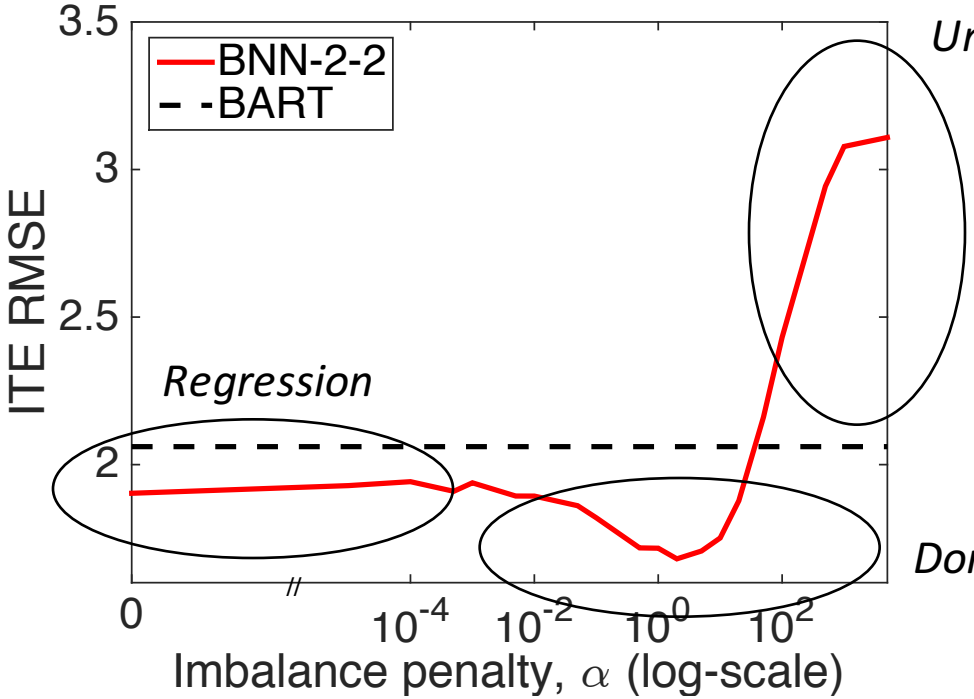
BNN-4-0



BNN-2-2



Results on IHDP



Objective

$$loss_F(\Phi, h) + \alpha \cdot disc(\Phi)$$

Upcoming work

Bounding and Minimizing Counterfactual Error

U. Shalit, F. Johansson, D. Sontag

https://arxiv.org/abs/1606.03976

Stronger theory, supporting different imbalance measures

Simpler, more flexible one-stage algorithm

Evaluation on real outcomes

Reliable Machine Learning Workshop
9:30 on Thursday!

Conclusions

We should care about causal and counterfactual inference in machine learning! Influence health care, education etc.

Many open questions: evaluation, confidence, hyperparameter selection