Domain Adaption
Representation Learning for CFI
Deep Learing Algorithm Outperforms State of the Art

} Abstract

# Introduction

Enforcing similarity between distributions of representations learned for populations with different interventions.

↳ representations of patients with Drug A or Drug B

↳ reduces variance from fitting a model on one distribution and applying it to another

Section 3: methods for learning such reps.

Section 4: "minimize upper bound on regret term in CF regime

In obs studies we have no control over mechanism that chooses actions nor feedback.

How can we learn from data, given that we don't know beforehand the particulars of determening actions which course of action would have better outcomes?

Domain Adaptation + Rep. Learning + Methods from Paper

1. Formulate problem of CFI as domain adaptation problem (covariate shift)
2. Derive families of representation algorithms for CFI
   - linear rep & var selection
   - deep learning of reps
3. Show that learning balanced reps. between treated & control populations leads to better counterfactuals (this is better than reweighting)

Intuitively, representations that reduce the discrepancy between the treated and control populations prevent the learner from using "unreliable" aspects of the data when trying to generalize from the factual to counterfactual domains. For example, if in our sample almost no men ever received medication A, inferring how men would react to medication A is highly prone to error and a more conservative use of the gender feature might be warranted.

# Problem Setup

For a context $x$ (e.g. patient) and for each potential intervention $t \in T$, let $Y_t(x) \in Y$ be the potential outcome of $x$.

Fundamental Problem: only one potential outcome is observed for $x$ → BANDIT FEEDBACK

$T = \{0, 1\}$ — 0: "control", 1: "treated"

- Individualized Treatment Effect (ITE)
  $Y_1(x) - Y_0(x)$
- Average Treatment Effect (ATE)
  $\mathbb{E}_{x \sim p(x)}[ITE(x)]$ for pop. with distribution $p(x)$
- factual outcome $y^F(x)$
- counterfactual outcome $y^{CF}(x)$

# We can't know ITE, but we can estimate it by direct modelling

Given $n$ samples $\{(x_i, t_i, y_i^F)\}_{i=1}^n$ (FACTUAL OUTCOME)

$$y_i^F = t_i \cdot Y_1(x_i) + (1-t_i) \cdot Y_0(x_i)$$

Learn a function $h : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$

$$h(x_i, t_i) \approx y_i^F$$

$$\hat{ITE}(x_i) = \begin{cases} y_i^F - h(x_i, 1-t_i), & t_i = 1 \\ h(x_i, 1-t_i) - y_i^F, & t_i = 0 \end{cases}$$

observed sample $\to$ $\hat{P}^F = \{(x_i, t_i)\}_{i=1}^n \sim P^F$

inferring $\to$ $\hat{P}^{CF} = \{(x_i, 1-t_i)\}_{i=1}^n \sim P^{CF}$

$P^F \neq P^{CF}$

empirical F/CF distributions

"feature dist. from training ≠ test"

Covariate Shift

$$P^F(x,t) = P(x) \cdot P(t/x)$$

$$P^{CF}(x,t) = P(x) \cdot \underbrace{P(\neg t/x)}_{\text{UNKNOWN}}$$

# Balancing Counterfactual Regression

Learned estimator $h$ must generalize from the factual distribution to the CF one.

## Algorithm

1. Input: $X, T, Y^F$; $\mathcal{H}, \mathcal{N}$; $\alpha, \gamma, \lambda$
2. $\Phi^*, g^* = \underset{\Phi \in \mathcal{N},\, g \in \mathcal{H}}{\arg\min}\ B_{\mathcal{H},\alpha,\gamma}(\Phi, g)$
3. $h^* = \underset{h \in \mathcal{H}}{\arg\min}\ \frac{1}{n} \sum_{i=1}^{n} \left(h(\Phi, t_i) - y_i^F\right)^2 + \lambda \|h\|_{\mathcal{H}}$
4. Output: $h^*, \Phi^*$

Learn rep $\Phi : \mathcal{X} \to \mathbb{R}^d$ (with deep NN) and func. $h : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}$ that trade off:

(1) low-error prediction of observed factual rep.

(2) " " " unobserved CF " from relevant factual outcomes

(3) balanced pop. distributions

Error minimization over training set ⇒ low-error prediction (1)
Penalty that encourages CF predictions close to nearest obs. outcome (2)
Minimizing discrepancy distance (3) FOR HYPO. SPACE $\mathcal{H}$, DISCREPANCY DISTANCE "$\text{disc}_{\mathcal{H}}$".

### 3.2. Deep neural networks

Deep neural networks have been shown to successfully learn good representations of high-dimensional data in many tasks (Bengio et al., 2013). Here we show that they can be used for counterfactual inference and, crucially, for accommodating imbalance penalties. We propose a modification of the standard feed-forward architecture with fully connected layers, see Figure 2. The first $d_r$ hidden layers are used to learn a representation $\Phi(x)$ of the input $x$. The output of the $d_r$:th layer is used to calculate the discrepancy $\text{disc}_{\mathcal{H}}(P_\Phi^F, P_\Phi^{CF})$. The $d_o$ layers following the first $d_r$ layers take as additional input the treatment assignment $t_i$ and generate a prediction $h([\Phi(x_i), t_i])$ of the outcome.

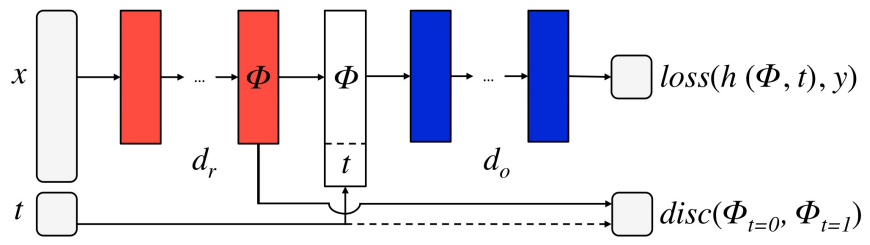


*Figure 2.* Neural network architecture.

For all fixed reps. $\Phi$ (a neural net), we have a bound on the relative error for a ridge regression model fit on the factual outcomes and evaluated on the counterfactual.
So we minimize the bound over $\Phi$ (Algorithm 1)

RM

SProp, with a small $l_2$ weight decay, $\lambda = 10^{-3}$. We evaluate two architectures. BNN-4-0 consists of 4 ReLU representation-only layers and a single linear output layer, $d_r = 4, d_o = 0$. BNN-2-2 consists of 2 ReLU representation-only layers, 2 ReLU output layers after the treatment has been added, and a single linear output layer, $d_r = 2, d_o = 2$, see Figure 2. For the IHDP data we use layers of 25 hidden units each. For the News data representation layers have 400 units and output layers 200 units. The nearest neighbor term, see Section 3, did not improve empirical performance, and was omitted for the BNN models. For the neural network models, the hypothesis and the representation were fit jointly.

We include several different linear models in our compari-

Bayesian Additive Regression Trees (BART) (Chipman et al., 2010) is a non-linear regression model which has been used successfully for counterfactual inference in the past (Hill, 2011). We compare our results to BART using the implementation provided in the BayesTree R-package (Chipman & McCulloch, 2016). Like (Hill, 2011), we do not attempt to tune the parameters, but use the default. Finally, we include a standard feed-forward neural network, trained with 4 hidden layers, to predict the factual outcome based on $X$ and $t$, without a penalty for imbalance. We refer to this as NN-4.

### 6.1. Simulation based on real data – IHDP

Hill (2011) introduced a semi-simulated dataset based on the Infant Health and Development Program (IHDP). The IHDP data has covariates from a real randomized experiment, studying the effect of high-quality child care and home visits on future cognitive test scores. The experiment proposed by Hill (2011) uses a simulated outcome and artificially introduces imbalance between treated and control subjects by removing a subset of the treated population. In total, the dataset consists of 747 subjects (139 treated, 608 control), each represented by 25 covariates measuring properties of the child and their mother. For details, see Hill (2011). We run 100 repeated experiments for hyperparameter selection and 1000 for evaluation, all with the log-linear response surface implemented as setting "A" in the NPCI package (Dorie, 2016).

### 6.3. Results

The results of the IHDP and News experiments are presented in Table 1 and Table 2 respectively. We see that, in general, the non-linear methods perform better in terms of individual prediction (ITE, PEHE). Further, we see that our proposed balancing neural network BNN-2-2 performs the best on both datasets in terms of estimating the ITE and PEHE, and is competitive on average treatment effect, ATE. Particularly noteworthy is the comparison with the network without balance penalty, NN-4. These results indicate that our proposed regularization can help avoid overfitting the representation to the factual outcome. Figure 4 plots the performance of BNN-2-2 for various imbalance penalties $\alpha$. The valley in the region $\alpha = 1$, and the fact that we don't experience a loss in performance for smaller values of $\alpha$, show that the penalizing imbalance in the representation $\Phi$ has the desired effect.

For the linear methods, we see that the two variable selection approaches, our proposed BLR method and LASSO + RIDGE, work the best in terms of estimating ITE. We would

On News, BLR and LASSO + RIDGE perform equally well yet again, although this time with qualitatively different results, as they do not select the same variables. Interestingly, BNN-4-0, BLR and LASSO + RIDGE all perform better on News than the standard neural network, NN-4. The performance of BART on News is likely hurt by the dimensionality of the dataset, and could improve with hyperparameter tuning.

## 7. Conclusion

As machine learning is becoming a major tool for researchers and policy makers across different fields such as healthcare and economics, causal inference becomes a crucial issue for the practice of machine learning. In this paper we focus on counterfactual inference, which is a widely applicable special case of causal inference. We cast counterfactual inference as a type of domain adaptation problem, and derive a novel way of learning representations suited for this problem.

Our models rely on a novel type of regularization criteria: learning *balanced representations*, representations which have similar distributions among the treated and untreated populations. We show that trading off a balancing criterion with standard data fitting and regularization terms is both practically and theoretically prudent.

Open questions which remain are how to generalize this method for cases where more than one treatment is in question, deriving better optimization algorithms and using richer discrepancy measures.