Linked Causal Variational Autoencoder for Inferring Paired Spillover Effects

ABSTRACT

Modeling spillover effects from observational data is an important problem in economics, business, and other fields of research. It helps us infer the causality between two seemingly unrelated set of events. For example, if consumer spending in the United States declines, it has spillover effects on economies that depend on the U.S. as their largest export market. In this paper, we aim to infer the causation that results in spillover effects between pairs of entities (or units); we call this effect as *paired spillover*. To achieve this, we leverage the recent developments in variational inference and deep learning techniques to propose a generative model called Linked Causal Variational Autoencoder (LCVA). Similar to variational autoencoders (VAE), LCVA incorporates an encoder neural network to learn the latent attributes and a decoder network to reconstruct the inputs. However, unlike VAE, LCVA treats the *latent attributes as confounders that are assumed to affect both the treatment and the outcome of units*. Specifically, given a pair of units u and $\bar{u}$, their individual treatment and outcomes, the encoder network of LCVA samples the confounders by conditioning on the observed covariates of u, the treatments of both u and $\bar{u}$ and the outcome of u. Once inferred, the latent attributes (or confounders) of u captures the spillover effect of $\bar{u}$ on u. Using a network of users from job training dataset (LaLonde (1986)) and co-purchase dataset from Amazon e-commerce domain, we show that LCVA is significantly more robust than existing methods in capturing spillover effects.

KEYWORDS

causal inference; spillover effect; variational inference; autoencoder; deep learning

ACM Reference Format:
. 2018. Linked Causal Variational Autoencoder for Inferring Paired Spillover Effects. In *Proceedings of ACM conference, , 2018 (CIKM18),* 4 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

Interference in causal inference is when the outcome of a unit is not only influenced by its own treatment and covariates but also by those of other units. In economics, this phenomenon is known as the *spillover effect*. Understanding spillover effects is extremely important to answer questions such as: Will eradicating pests from one farm cause them to move to nearby farms? Will the popularity of one product drive the sales of another product?

CIKM18, 2018,

ACM ISBN . . . $15.00
https://doi.org/10.475/123_4

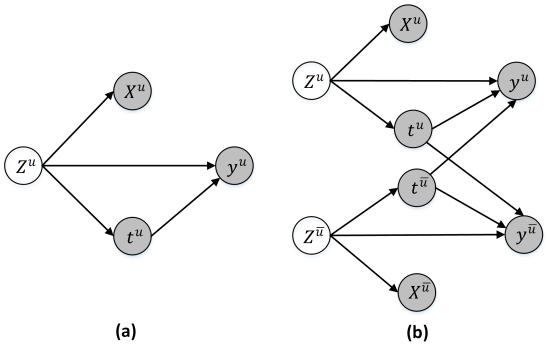

Figure 1: Graphical structure of (a) conventional causal-effect model and (b) linked causal-effect model. Here, u and $\bar{u}$ are a pair of units, X^* are the observed covariates (or proxies) of units, t^* is the treatment and y^* is the outcome. The superscript $*$ indicates either u or $\bar{u}$.

Will the introduction of genetically modified crops result in the contamination of neighboring organic crops? In this paper, our goal is to estimate the causal effect of a certain treatment (e.g. positive review) on a specific outcome (e.g. product sales) when there exist spillover effects between pairs of entities. For instance, positive reviews for XBOX consoles could result in increased sales of other XBOX accessories such as controllers and games. We term this effect as *paired spillover*.

Research on causal Machine Learning has gained significant attention in recent years. This can be attributed to their empirical success in estimation of unknown functions without strong model specification. For instance, in [1], the authors show that direct regression adjustment with Bayesian Additive Regression Trees (BART) leads to promising estimation of individual treatment effects. [2] propose a model for counterfactual inference by bringing together ideas from representation learning and domain adaptation. [13] and [8] utilize deep learning techniques to infer individual treatment effects and counterfactual outcome. Despite their novel approach, none of these studies account for spillover effects.

To overcome the aforementioned problem, we leverage the latest developments in variational deep learning techniques [3, 8] to propose a model called Linked Causal Variational Autoencoder (LCVA). The framework of LCVA is based on variational autoencoder (VAE) [3], which incorporates an encoder neural network to learn the latent attributes and a decoder network to reconstruct the inputs. Nonetheless, unlike VAE, LCVA adopts the inferential process of a causal variant of VAE called *causal effect variational autoencoder* (CEVAE) [8]. The graphical structure shown in Figure 1 (a) explains the generative principle of CEVAE. Here, X is a set of covariates of a unit u, t is the treatment, y is the outcome and Z is the confounder. CEVAE treats the confounder as a latent variable and conditions the outcome and treatment on the hidden confounder. We propose an extended framework in Figure 1 (b) where u and $\bar{u}$ are two units

and the spillover effect is modeled by allowing the treatment of unit u (i.e., t^u) to influence the outcome of unit $\bar{u}$ (i.e., $y^{\bar{u}}$) and vice versa. Our objective is to learn the confounders Z^u and $Z^{\bar{u}}$. A common practice of inferring such counfounders is to use *proxy variables* [9, 12]. For example, in the study of causal effect of job training on annual income, we cannot measure every attribute that influences the earnings of an individual, but we might be able to get a proxy for it through a set of accessible variables such as zip code and job type. There are several ways to use these proxies to estimate Z. Louizos et al. [8] showed that one of the effective ways is to directly condition the proxy on Z and infer Z using approximate maximum-likelihood based methods. Therefore, we adopt the same technique by allowing the covariates X^u and $X^{\bar{u}}$ to act as proxies of confounders Z^u and $Z^{\bar{u}}$, respectively.

To evaluate the proposed model, it is important to consider datasets where the outcome of units is influenced by some form of spillover effect. Ideally, the dataset needs to have the following properties: (1) network information in the form of links between units and (2) the counterfactual outcome of individual units. Unfortunately, most existing datasets do not contain both these properties together. Therefore, we modify the following real word datasets by filling-in the missing information: (a) the job training dataset [6, 14] and (b) the co-purchase dataset from Amazon e-commerce domain [10]. In particular, the job training dataset does not include any network information; consequently, similar to [7], we create a K-NN graph based on the covariates of units to connect similar individuals. This can be justified by the theory of Homophily [11] which states that *birds of a feather flock together*. Contrary to the job training dataset, the co-purchase dataset from Amazon does contain the network information, which specifically states whether an item is a substitute or a complement of another item [10]. Nonetheless, it does not have the counterfactual outcome. In our case, the counterfactual is the sales of a product if it had no reviews. We synthesize this using a matching technique that was introduced by Kun et al. [5]. The major contributions of this paper are detailed as follows:

- We propose a model called linked causal variational autoencoder (LCVA) that captures the spillover effect between pairs of units. Specifically, given a pair of units u and $\bar{u}$, their individual treatment and outcomes, the encoder network of LCVA samples the confounders by conditioning on the observed covariates of u, the treatments of both u and $\bar{u}$ and the outcome of u.
- We introduce two datasets by augmenting the job training dataset [6, 14] with synthesized network information and augumenting Aamzon dataset [10] with counterfactual outcomes.
- Using a rigorous series of experiments, we show that LCVA is extremely effective in capturing spillover effects between units. It also beats existing methods on metrics such as Average Treatment Effect, PEHE, and Policy Risk across all datasets.

To the best of our knowledge, this is the very first deep variational inference framework that is specifically designed to infer the causality of spillover effects between pairs of units.

2 THE PROPOSED LCVA MODEL

As explained in the previous section, our objective is to infer the latent confounders Z^*, where $*$ indicates u or $\bar{u}$. To achieve this, we make the following assumption: the latent confounder z^* can be sampled from the observed variables x^*, t^*, y^* and the treatment

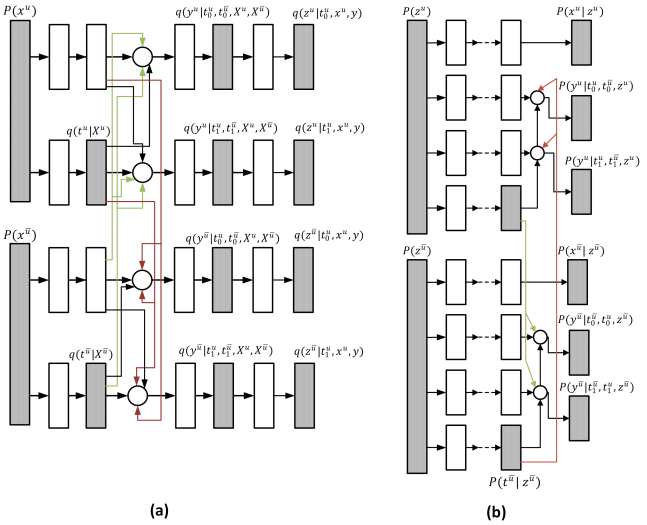

Figure 2: The architecture of Linked Causal Variational Autoencoder (LCVA) model; (a) is the encoder part of the neural network and (b) is the decoder. The red lines indicate the effect of unit u's covariates and treatment on the network of $\bar{u}$ and the red lines indicate the same from $\bar{u}$ to u.

$t^{\hat{*}}$, where $\hat{*}$ indicates $\bar{u}$ if $* = u$ and u if $* = \bar{u}$, x indicates the covariates of a single unit and z indicates the confounder of a single unit. z^* is inferred using the proposed linked causal variational autoencoder (LCVA) that is depicted in Figure 2 . Here, the white nodes correspond to parametrized deterministic neural network, the grey nodes correspond to drawing samples from the respective distribution and the white circles correspond to switching path according to the treatment t. Learning the latent variable z is typical of any variational inference technique, where the objective is to optimize the KL-divergence between the true posterior $p(z^*|\cdot)$ and the variational distribution $q(z^*|\cdot)$. The sampling of $q(z^*|\cdot)$ is similar to the recently proposed CEVAE [8]; however, unlike CEVAE, our goal is to capture the paired spillover effect. Therefore, in Figure $q(z^*|\cdot)$ (where $* = u$ or $\bar{u}$) is sampled not only based on its individual covariates x^*, treatments t^*, and factual outcome y^*, but also on $t^{\bar{u}}$ and $x^{\bar{u}}$. The green and yellow lines in Figure 2 (a) signify this dependency.

2.1 Parameter Inference

We begin by defining the evidence lower bound (ELBO) of LCVA for unit u as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_i^u|x_i^u,x_i^{\bar{u}},t_i^u,t_i^{\bar{u}},y_i^u)} \Big[log P_\theta(x_i^u|z_i^u) + log\, P_\theta(t_i^u|z_i^u) + log\, P_\theta(y_i^u|t_i^u,t_i^{\bar{u}},z_i^u) - log\, q_\phi(z_i^u,t_i^u,t_i^{\bar{u}},y^u|x_i^u,x_i^{\bar{u}}) \Big] \quad (1)$$

where N is the number of units, $\mathbb{E}_{q(z_i^u|\cdot)}$ is the expectation w.r.t distribution q and ϕ, θ are the weights of encoder and decoder network respectively. The ELBO of unit $\bar{u}$ remains similar to the above expression; hence, we do not exclusively derive the lower bound for $\bar{u}$. From the encoder's neural network (Figure 2(a)), one can observe

that the approximate posterior $q(z|\cdot)$ factorizes as follows:

$$\sum_{i=1}^{N} q_\phi(z_i^u|t_i^u, t_i^{\bar{u}}, x_i^u, x_i^{\bar{u}}, y^u) + log\, q_\phi(t_i^u|x_i^u) + log\, q_\phi(t_i^{\bar{u}}|x_i^{\bar{u}}) + log\, q_\phi(y_i^u|x_i^u, x_i^{\bar{u}}, t_i^u, t_i^{\bar{u}}) \quad (2)$$

we can obtain an unbiased estimate of the ELBO by sampling $z^u \sim q_\phi$ and use stochastic gradient descent to optimize it. However, we cannot trivially take gradients w.r.t ϕ. Therefore, we incorporate the *reparamaterization trick* [3], to sidestep this issue. $q(z^u)$ is then approximated by the following expression:

$$q(z_i^u|t_i^u, t_i^{\bar{u}}, x_i^u, x_i^{\bar{u}}, y^u) = \prod_{j=1}^{K} \mathcal{N}(\mu_{ij}, \sigma_{i,j}^2)$$

$$\boldsymbol{\mu_i} = t_i\,(\mu_{t=0,i}^u + \mu_{t=0,i}^{\bar{u}}) + (1-t_i)\,(\mu_{t=1,i}^u + \mu_{t=1,i}^{\bar{u}})$$

$$\boldsymbol{\sigma_i} = t_i\,(\sigma_{t=0,i}^u + \sigma_{t=0,i}^{\bar{u}}) + (1-t_i)\,(\sigma_{t=1,i}^u + \sigma_{t=1,i}^{\bar{u}}) \quad (3)$$

where K is the number of latent features, μ_t is the mean of units that received treatment t and σ_t indicates the same for variance. As explained earlier, since z^u is unobserved, in the inference network, the outcome y^u is influenced by both $t^{\bar{u}}$, $x^{\bar{u}}$ due to property of *common cause* [4]. However, once z^u is inferred, the outcome y^u simply depends on the sampled treatments t^u and $t^{\bar{u}}$. Therefore, in the decoder part (Figure 2(b)), the variables can be generated as follows:

$$p(\boldsymbol{x_i^u}|z_i^u) = \prod_{j=1}^{N} p(x_{i,j}^u|\boldsymbol{z_i}) \quad (4)$$

$$p(\boldsymbol{t_i^u}|z_i^u) = Bernoulli(\sigma(g_1(z_i))) \quad (5)$$

$$p(y_i^u|t_i^u, t_i^{\bar{u}}\boldsymbol{z_i^u}) = \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i)\;\; \hat{\mu}_i = t_i^u g_2(z_i^u) + (1-t_i^u) g_3(z_i^u) \quad (6)$$

where $p(x_{i,j}^u|\boldsymbol{z_i})$ is a gaussian distribution, $\sigma(\cdot)$ is the logistic function, and $g_{\{\cdot\}}$ is a neural network parameterized by weights θ.

3 EXPERIMENTS

In this section, we intend to answer the following question: how accurate can the proposed model be in terms of inferring treatment effects?. To see the effectiveness of the proposed model, we compare it with a state-of-the-art model and four other baselines that are widely regarded as classical methods in causal inference. For each model, we carried out a grid search to decide the hyperparameters.

OLS-1. A linear regression model $f : [X^u, t^u, t^v] \rightarrow y^u$ is trained with the treatment of linked units considered. We can infer the counterfactual outcomes by applying f on $[X^u, 1-t^u, t^v]$ for all pair of units $(u, v) \in \mathcal{E}$.

OLS-2. Given $(u, v) \in \mathcal{E}$, we train two linear regression models $f_1 : [X^u, t^v] \rightarrow y^u, \forall u : t^u = 1$ and $f_0 : [X^u, t^v] \rightarrow y^u, \forall u : t^u = 0$. Then the counterfactual outcomes can be inferred by $\hat{y}_1^u = f_1(X^u, t^v), \forall u : t^u = 0$ and $\hat{y}_0^u = f_0(X^u, t^v), \forall u : t^u = 1$.

Random forest. This model is the same as OLS-1 except the function f is replaced by a random forest.

Causal forest [15]. This ensemble model consists of a set of causal trees. Each causal tree splits the original feature space into leaves and considers the treatments and outcomes of units in a leaf to come from a randomized set of experiments.

CEVAE [8]. Similar to LCVA, CEVAE learns confounders Z^u for each unit but does not take the spillover effect into consideration.

Table 1: Statistics of Job training dataset and co-purchase dataset of positive and negative reviews from Amazon.

Name	#Treated	#Control	#Pairs	#Feature
Job	297	2915	3512	10
+ve Amazon	50K	10K	96132	300
-ve Amazon	20K	5K	28136	300

3.1 Dataset and Evaluation Metrics

Job training dataset: This dataset is used to study the treatment effect of job training on the earning in the year of 1978. In order to de-randomized the data, following [13], we keep the treatment group of LaLonde's study [6] ($\mathcal{U}_T$) and combine the control group of this study ($\mathcal{U}_C$) and that from the PSID [14] ($\mathcal{U}_{C1}$). Additionally, since this dataset does not include any network information, we create a K-NN graph based on the covariates of units to connect similar individuals. Table 1 shows the statistics of our dataset.

Evaluation metrics for job dataset: Due to lack of ground truth for the counterfactual outcomes, we use average treatment effect on the randomized trial subset and policy risk as evaluation metrics [13]. The fact that the LaLonde's study is a randomized trial enables us to get the approximated ground truth of the ATE for the subset $\mathcal{U}_T \cup \mathcal{U}_C$ by the naive estimator $ATE = \frac{1}{|\mathcal{U}_T|}\sum_{u\in\mathcal{U}_T} y_1^u - \frac{1}{|\mathcal{U}_C|}\sum_{u\in\mathcal{U}_C} y_0^u$. Then, the estimated ATE can be calculated as $\hat{ATE} = \frac{1}{|\mathcal{U}_T\cup\mathcal{U}_C|}\sum_{u\in\mathcal{U}_T\cup\mathcal{U}_C}(\hat{y}_1^u - \hat{y}_0^u)$, where y^u and $\hat{y}^u$ refer to the observed factual and estimated counterfactual outcome, respectively. For a treated (or controlled) unit, $\hat{y}_1^u = y_1^u$ ($\hat{y}_0^u = y_0^u$) and $\hat{y}_0^u$ ($\hat{y}_1^u$) is inferred by the models. We report the absolute difference between the estimated ATE and the approximated ground truth: $\epsilon_{ATE} = |ATE - \hat{ATE}|$. Moreover, following the evaluation methodology of [13], we also report the estimated policy risk ($\hat{pr}$) for the randomized trial subset which is defined as $\hat{pr} = 1 - (\mathbb{E}[\tilde{y}_1^u|t^u = 1, \pi^u = 1]p(\pi^u = 1) + \mathbb{E}[\tilde{y}_0^u|t^u = 0, \pi^u = 0]p(\pi = 0))$, where $\pi^u = \mathbf{1}(\hat{y}_1^u - \hat{y}_0^u)$ is the indicator function of whether the estimated individual treatment effect is positive and $\tilde{y}^u$ denotes the factual outcome that is scaled between $[0, 1]$. Intuitively, the weighted sum of the two expectations denotes the expected potential outcome.

Amazon dataset: For the Amazon dataset [10], we study the causal effect of positive (or negative) reviews on the sales of products. For our experiments, we choose the co-purchase data from the electronics category and divide the products (or units) into two groups (1) units that have more than three reviews and (2) units that have less than three reviews. The first group is considered as treated (i.e., t=1), while the second is the control group (i.e., t=0). Considering the fact that positive and negative reviews can affect the sales in different ways, we separate the units in treated group into two different datasets: (a) units with positive reviews (when average rating > 3) and (b) units with negative reviews (when average rating < 3). To each of these dataset, we add the units from control group to create the final dataset (Table 1). The features of each unit is created by feeding the review text to a doc2vec model to create a vector of 300 latent features per unit. Lastly, we overcome the lack of counterfactual outcomes using a matching technique [5] to synthesize counterfactuals. To be specific, for a product u, the counterfactual outcome is set as the observed sales of the most

Table 2: Results for the job training dataset.

Models	ϵ_{ATE}	Policy Risk
OLS1	492.51	0.87
OLS2	498.05	0.86
RF	4.05	0.84
CF	511.68	0.93
CEVAE	112.46	0.84
LCVA	55.63	**0.794**

Table 3: Results for Amazon dataset-positive reviews.

Models	ϵ_{ATE}	PEHE
OLS1	8.34	103.90
OLS2	7.99	92.27
RF	9.46	83.92
CF	13.49	153.35
CEVAE	8.39	55.31
LVAE	**1.037**	**13.107**

similar product with an opposite treatment status i.e., $y^u_{1-t^u} = y^v$, where $v = argmin_{v:t^v=1-t^u} ||X^v - X^u||^2_2$.

Evaluation metrics for Amazon dataset: We use the following metrics for evaluation: Precision in Estimation of Heterogeneous Effect (PEHE) and absolute error on Average Treatment Effect (ATE) as $PEHE = \frac{1}{N}\sum_u((y^u_1 - y^u_0) - (\hat{y}^u_1 - \hat{y}^u_0))^2$, $\epsilon_{ATE} = |ATE - \hat{ATE}|$, where $ATE = \frac{1}{N}\sum_u(y^u_1 - y^u_0)$ and $\hat{ATE} = \frac{1}{N}\sum_u(\hat{y}^u_1 - \hat{y}^u_0)$.

3.2 Results

Table 2 compares the performance of LCVA along with other baselines for job training dataset, and Tables 3 and 4 reports the same for co-purchase dataset from Amazon. Overall, our model consistently outperforms the baselines in almost all scenarios. This observation can be explained by the fact that although models such as OLS1, OLS2 and random forest can learn the spillover effect, they only do so by controlling the observable features. Unfortunately, these features are inadequate to represent all the confounding variables. In comparison, LCVA learns representation for confounders with information extracted not only from features but also from treatments and factual outcomes. Another interesting observation is that for the job training dataset (Table 2), RF performs better than our model in terms of ϵ_{ATE}. However, this scenario is different when it comes to Amazon dataset where LVAE is significantly better than RF on both positive and negative cases. A possible reason for this outcome could be attributed to the level (or intensity) of spillover effects in datasets. In job training dataset, we synthetically create the links between units using K-NN algorithm, while in Amazon dataset, the link relationship between products is naturally present. This in turn implies that spillover effects are much stronger in Amazon dataset due the co-purchase behavior. Finally, it is also important to note that LCVA achieves better estimation of treatment effects (or counterfactual outcomes) when compared to the state-of-the-art CEVAE. This is because our model is specifically designed to capture the spillover effect between linked units.

4 CONCLUSION

In this paper, we propose a model called linked causal variational autoencoder (LCVA) that captures the spillover effect between pairs of units. Specifically, given a pair of units u and $\bar{u}$, their individual treatment and outcomes, the encoder network of LCVA samples the confounders by conditioning on the observed covariates of u, the treatments of both u and $\bar{u}$ and the outcome of u. Using a network of users from job training dataset (LaLonde (1986)) and co-purchase dataset from Amazon e-commerce domain, we show that LCVA is significantly more robust than existing methods in capturing spillover effects.

Table 4: Results for Amazon dataset-negative reviews.

Models	ϵ_{ATE}	PEHE
OLS1	11.25	52.50
OLS2	2.7	57.76
RF	11.43	49.50
CF	9.43	55.32
CEVAE	7.64	43.72
LVAE	**1.218**	**13.107**

REFERENCES

[1] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* (2011).

[2] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International Conference on Machine Learning*.

[3] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[4] Daphne Koller and Nir Friedman. [n. d.]. *Probabilistic graphical models: principles and techniques*. MIT press.

[5] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. In *Proceedings of the 23rd ACM SIGKDD*.

[6] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* (1986).

[7] Jundong Li, Liang Wu, Dani Harsh, and Huan Liu. 2018. Unsupervised Personalized Feature Selection. In *Proceedings of The 32nd AAAI*.

[8] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*.

[9] Gangadharrao Soundaryarao Maddala and Kajal Lahiri. 1992. *Introduction to econometrics*. Macmillan New York.

[10] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD*.

[11] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* (2001).

[12] Mark R Montgomery, Michele Gragnolati, Kathleen A Burke, and Edmundo Paredes. 2000. Measuring living standards with proxy variables. *Demography* (2000).

[13] Uri Shalit, Fredrik Johansson, and David Sontag. 2016. Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint arXiv:1606.03976* (2016).

[14] Jeffrey A Smith and Petra E Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics* (2005).

[15] Stefan Wager and Susan Athey. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* (2017).