# Treatment Effect Estimation of the DragonNet under Overlap Violations

Marco van Veen
Supervisor(s): Stephan Bongers, Jesse Krijthe
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

**Abstract**

The large amounts of observational data available nowadays have sparked considerable interest in learning causal relations from such data using machine learning methods. One recent method for doing this, which provided promising results, is the DragonNet (Shi et al., 2019), which utilises neural networks in order to estimate average treatment effects in populations. The performance of the model, however, was not tested on datasets which contain low amounts of overlap between the treated and non-treated subpopulations, which makes it harder to accurately estimate treatment effects. Therefore, the goal of this paper is to investigate the performance of the DragonNet when used on datasets with (near) overlap violations. This has been done by looking at the mean absolute errors and variances of the estimated treatment effects and comparing these to other models. The results showed that the performance of the DragonNet becomes significantly worse compared to other models when large portions of the population suffer from low overlap. Additionally, the variance of the results also increases in these cases, making the results less reliable. From the obtained results, it can be concluded that it is best to choose another model for treatment effect estimation if relatively large amounts of overlap violations are suspected.

# 1 Introduction

Understanding the causal effects of actions is a key task in many fields of research. Examples of such domains are medicine, where the effect of some treatment is being investigated, or economics, where the effects of various policies on the economy may be of interest. Randomised controlled trials can be used in order to properly estimate causal effects (Shadish et al., 2001), but they are not always feasible or ethical. Therefore, machine learning methods have become increasingly popular for estimating causal effects from the large amounts of observational data that are available nowadays.

Traditional machine learning methods mainly focus on predictive or descriptive tasks. In order to perform these tasks, associations between actions and outcomes within the datasets are used in order to predict new outcomes. Such associations, though, may (often) not reflect the true causations within the data. A correlation between two variables may be due to both of them having a common cause which drives their values. Such a variable that affects both the actions and the outcomes is known as a confounding variable. For example, from some observational data it may be concluded that patients receiving some treatment tend to have a longer life expectancy. However, it could be the case that only wealthy individuals were able to afford the treatment. Such wealthier people are also more likely to live a healthier lifestyle in general, leading to a higher life expectancy. Therefore, in order to properly study the effect of the treatment on the life expectancy, confounding variables such as wealth must be taken into account.

Guo et al. (2020) have provided a comprehensive survey on the current state-of-the-art models for estimating causal effects which take into account such confounding variables. Many of these models seem to make use of neural networks or tree structures. Within the group of neural networks for causal effect estimation, a number of them have focused on learning alternative representations of the confounding variables, which may lead to better causal effect estimations than using the original features in the datasets. One such method which has achieved good causal estimation results is the TARnet (Shalit et al., 2017).

A recent extension to the TARnet, which was not discussed in the survey, is the DragonNet (Shi et al., 2019). DragonNet tries to improve upon the original TARnet by including the estimation of the propensity score $P[T = 1|\mathbf{x}]$ within the model, which is the probability of receiving treatment $T$ given a set of features. By including this estimation within the model, the goal is for only the confounding variables, which affect both the treatment and outcome, to be used while discarding the other irrelevant features. Additionally, a targeted regularisation procedure is applied to the model which will transform the estimator of the average treatment effect into a robust and efficient estimator. These desirable properties are expected to improve the finite-sample performance of the model.

In order for the DragonNet to be able to estimate the treatment effects, two important assumptions have to be made. The first assumption says that there are no hidden confounders, i.e., the data contains all possible confounding variables. Such hidden confounders which affect the outcomes can lead to incorrect conclusions about the treatment effects. The second assumption is that of overlap, which says that $0 < P[T = 1|\mathbf{x}] < 1$ must hold, where $T = 1$ represents a binary treatment assignment. This is required, since it is impossible to estimate the treatment effect for some group of individuals with features $\mathbf{x}$ if either none or all of them have received treatment. This assumption is especially of interest when looking at the DragonNet, as they extended the TARnet model by including propensity score estimation within the model and also used propensity scores in the targeted regularisation procedure.

However, the effects of propensity scores very close to 0 or 1 have not been thoroughly investigated in the DragonNet paper. Datasets containing large amounts of overlap violations have been ignored when testing the model. Additionally, for the datasets which were used for testing, individual points with predicted propensity scores close to 0 or 1 were discarded when estimating the final treatment effect. Understanding the behaviour of the model under such overlap violating conditions is still important, since the assumption does not always hold in real-world data. It can provide insights into whether the DragonNet seems like a suitable model for causal effect estimation in real-world applications, or whether other models may prove to be better alternatives.

Therefore, the main goal of this paper will be to answer the following question: "What is the effect of overlap violations within datasets on the estimation performance of the DragonNet model?". In order to answer to this question, the DragonNet model will be tested on synthetic data with various amounts of overlap violations. Additionally, it will be tested on some synthetic real-world benchmark datasets which contain overlap violations and the results will be compared to those from a number of alternative models.

The remainder of this paper will be structured as follows. Section 2 provides a more detailed explanation of the DragonNet model and the methods used to analyse its performance. After that, Section 3 describes the specific experiments performed in more detail. Next, Section 4 presents the results from performing the experiments described in the previous section and draws some conclusions based on these results. Section 5 then continues by reflecting on the obtained results in a broader context and comparing it to results obtained from previous experiments. In Section 6, the ethical aspects of this research and the reproducibility of the experiments are discussed. Finally, Section 7 concludes this paper by providing a brief summary of work and presenting the main conclusions. Additionally, it discusses potential topics for further research.

# 2 Methodology

In this section the DragonNet model and the experimental methods which were used to understand the DragonNet model's behaviour under overlap violations will be described. The first subsection will provide a brief overview of the DragonNet model. After that, the second subsection will focus on how structural causal models have been used to generate synthetic data and how this data is then used for analysing the DragonNet model.

## 2.1 DragonNet

The goal of the DragonNet model (Shi et al., 2019) is to estimate the average treatment effect (ATE) of some binary treatment from observational data using neural networks. It aims to do this by first predicting what the outcomes of all individuals would be under treatment and no treatment, and then subtracting and averaging these two outcomes to estimate the ATE. The model tries to exploit Theorem 3 from Rosenbaum and Rubin (1983), which essentially states that, when estimating the ATE, it is sufficient to only use the features from the data which are required for predicting the propensity scores. Therefore, the model should discard other, irrelevant features, as they are simply considered as noise, which does not help with estimating the treatment effect. It tries to achieve this by combining the estimation of the outcomes and the propensity scores in the objective function, such that the neural network will be trained to discard the irrelevant features.

The architecture of the DragonNet can be seen in Figure 1. It takes an input X and puts it through a three-layered neural network to obtain the new, shared representation Z. This shared representation is then used to estimate the propensity scores $\hat{g}$ and the conditional outcomes under treatment ($\hat{Q}(1,\cdot)$) and under no treatment ($\hat{Q}(0,\cdot)$), all at once. This architecture is very similar to the TARnet (Shalit et al., 2017), but with the additional $\hat{g}$ head.

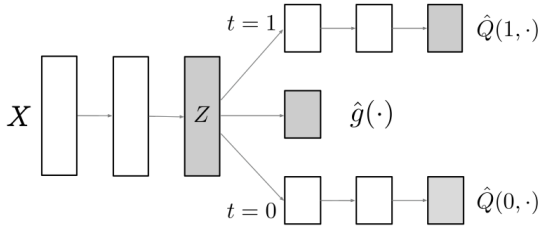



Figure 1: The DragonNet Architecture. Transform input X to an alternative representation Z using neural networks and train to predict outcomes with ($\hat{Q}(1,\cdot)$) and without ($\hat{Q}(0,\cdot)$) treatment, and the propensity scores $\hat{g}$ at once. The grey layers represent layers which produce some form of output (Z, $\hat{Q}(\cdot,\cdot)$, or $\hat{g}$). From: Shi, C., Blei, D., and Veitch, V. (2019). "Adapting neural networks for the estimation of treatment effects".

Aside from adding the propensity score head $\hat{g}$ to the TARnet model, Shi et al. (2019) added an additional improvement called "targeted regularisation". This procedure essentially modifies the outcome model and objective function using non-parametric estimation theory in order to obtain desirable asymptotic properties for the ATE estimator. These properties are robustness in the double machine-learning sense and efficiency of the ATE estimator. Double machine learning has been introduced by Chernozhukov et al. (2016) and Chernozhukov et al. (2017) and states that, if certain equations are satisfied, the ATE

estimator will converge to the true value at a fast rate as long as $\hat{g}$ and $\hat{Q}(\cdot, \cdot)$ converge. More in-depth details about the targeted regularisation procedure can be found in Shi et al. (2019).

An important detail about the targeted regularisation is that the modification to the outcome model and objective function contains a term which depends on the estimated propensity scores $\hat{g}(\cdot)$ and treatment assignments $t_i$, namely $\frac{t_i}{\hat{g}(\cdot)} - \frac{1-t_i}{1-\hat{g}(\cdot)}$. Due to the usage of estimated propensity scores in these denominators, the expectation is that the model is sensitive to the extreme propensity scores in datasets with low overlap. Therefore, even though the model showed very promising results for datasets which have no (near) overlap violations (Shi et al., 2019), it may perform significantly worse on datasets which have low overlap.

## 2.2 Structural Causal Models

In order to generate data with desired causal relations, the structural causal models (SCMs) introduced by Pearl (2009) can be used. SCMs allow for the mathematical formulation of the causal relations between variables through the use of *structural equations*. The general form of these equations used for the experiments can be seen in Equation 1.

$$\begin{aligned} X &= f_X(\epsilon_X) \\ T &= f_T(X, \epsilon_T) \\ Y &= f_Y(X, T, \epsilon_Y) \end{aligned} \tag{1}$$

where X denotes confounding variables, T denotes the assigned treatments (either 0 or 1), and Y denotes the outcomes. The disturbance terms $\epsilon_X$, $\epsilon_T$, and $\epsilon_Y$ are assumed to be mutually independent and arbitrarily distributed. These terms represent the effects of unobserved, exogenous variables (Pearl, 1995). Additionally, since overlap violations are the main interest in this work, this model operates in a "no hidden confounding" setting as the disturbance terms are mutually independent and thus X represents all possible confounders.

Datasets generated using the above formulation, with varying propensity scores, will be used to analyse the ATE estimator of the model. Even though synthetic data does not capture the complexity of actual real-world problems, experiments performed using such data can still provide valuable information about the behaviour of the model. This is due to the fact that it is possible to repeatedly generate the same data using slightly different settings, such as different propensity scores, which then allows for analysing the effects of overlap violations in an isolated setting. The analysis will be performed in a number of ways.

First, mean absolute errors (MAEs) of the estimator will be obtained for datasets which contain different amounts of overlap violations. As propensity scores represent the probabilities of receiving treatment given a set of features, the very low or high treatment probabilities can be assigned to different sub-populations which contain individuals with similar features. Therefore, the MAEs will be analysed by varying both the percentage of the population that suffers from bad propensity scores and the value of the propensity scores for these sub-populations.

Next, the convergence and rate of convergence of the ATE estimator will be obtained under different propensity scores by varying the sample sizes of the datasets. The ATE estimator of the DragonNet model is consistent and should converge to the true ATE value quickly due to the targeted regularisation procedure within the model. As the model was

originally not investigated under (near) overlap violations, performing these experiments can provide a better understanding of how quickly these desirable properties break down.

Finally, the variance of the results will also be analysed by comparing box-plots of the obtained MAEs for different configurations. A high variance makes it hard to trust the results of the model when ran on a specific dataset. Therefore, understanding the effects of bad propensity scores on the variance can help with identifying how useful the results of the model are in such cases.

# 3 Experiments

This section will provide details for the experiments which were performed for analysing the DragonNet model. The first subsection will present the specific settings used for the models during the experiments. Next, the experiments using synthetic data are explained. Finally, the last subsection outlines how experiments using more realistic, semi-synthetic datasets were performed.

## 3.1 Model Setup

The DragonNet model will be run using the same settings as described in Shi et al. (2019). The hyper-parameters $\alpha$ and $\beta$ in Equations 2.2 and 3.2 in Shi et al. (2019) are set to 1. The sizes of the hidden layers for the shared representation are 200 and for the conditional outcomes 100.

The DragonNet model without data trimming will be compared to the DragonNet model with data trimming and the TARnet model. The DragonNet model with data trimming will discard any data point with an estimated propensity score outside of [0.05, 0.95] when calculating the ATE. The TARnet model from Shalit et al. (2017) has the same implementation and settings as the DragonNet, but with the propensity score head removed from the model and without the targeted regularisation procedure.

All the results are obtained using 200 replications of the datasets for each different configuration of parameters. Additionally, the data is not split into train and test sets, but all of the data is used for both training the model and estimating the ATE. This approach is also used in Shi et al. (2019). This can be considered a valid approach in this case, as the main goal of the model is to simply estimate the ATE value for the specific dataset at hand without the goal of generalising the model to new datapoints. Overfitting is, therefore, not as much of an issue in this setting. However, it is possible to overfit on the observed data to some degree, as the model should still generalise to newly observed outcomes from the same populations used in the datasets.

## 3.2 Synthetic Data Experiments

The synthetic data is created using an SCM which allows for splitting the population in 2 parts; one which has a perfectly balanced probability of treatment (0.5) and one which has low probabilities of treatment. Three covariates $X_i \sim U(0,1)$ are generated. Then, $T \sim Ber(0.5)$ if $X_3$ is above some threshold and $T \sim Ber(p)$ if $X_3$ is below the threshold, where $p$ is some (low) probability of receiving treatment in order to generate subpopulations with bad propensity scores. The size of the subpopulation suffering from these bad propensity scores can be varied depending on the threshold dividing $X_3$. Finally, the outcome model is $Y = 1 + T + X_1 + 2X_2 + 0.5X_3 + \epsilon_Y$, where $\epsilon_Y \sim N(0,1)$ represents the effects of exogenous

random variables which are not of interest. The causal relations between the variables are visualised in Figure 2.

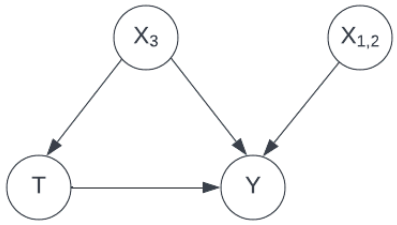



Figure 2: Graph shows the causal relations between the variables. $X_3$ has an effect on the probability of receiving binary treatment T (propensity score) and on the outcome Y, while $X_1$ and $X_2$ only affect Y. The arrow from T to Y indicates the treatment effect.

The number of covariates used here, namely three, was arbitrarily chosen. Any small number of covariates, such as two, should also work in this simplistic setting and lead to similar results in the end. The default model settings described in Section 3.1 are used for all experiments.

### 3.3 Semi-Synthetic Data Experiments

As ground truth values of treatment effects are generally not available for real-world data, synthetic real-world datasets will be used in order to test the model in a more realistic setting compared to the synthetic data setting from above. A widely used semi-synthetic dataset is the IHDP dataset originally provided by Hill (2011), which is based on the Infant Health and Development Program (IHDP). The dataset consists of 747 datapoints, of which 139 are treated, and has 25 covariates. This dataset has the underlying treatment effects used during the data generation available and the covariates are representative of those from a real observational study, which allows for testing the DragonNet in a more realistic setting compared to the simple synthetic one.

One issue with such datasets is the fact that it is harder to test the model for different propensity scores, as it is based on observed real-world data. Shalit et al. (2017) found a way to still create artificial imbalances between the control and treated groups in order to still be able to test their model in such imbalanced scenarios. They did this by randomly removing observations with the highest estimated propensity scores from the control group with a probability of $q$ and removing random control observations with a probability of $1-q$. This was done until 400 observations remained in each IHDP sample dataset.

A similar approach will be used here in order to test the DragonNet under extreme propensity scores with the IHDP datasets. The degree of overlap violations can be modified by choosing different values of $q$ when randomly removing observations. A higher $q$ will cause an increased imbalance between the treated and control groups, thus leading to a lower propensity score. This will allow for performing similar tests as before by calculating the MAE and variance of the estimator in these more realistic cases.

The IHDP samples used here are 200 samples randomly taken from the 1000 generated replications used by Shalit et al. (2017) [1]. These samples were generated using the NPCI

[1] All datasets can be found at: https://www.fredjo.com/

package (Dorie, 2016) with setting "A", which corresponds to setting (or "response surface") "B" in Hill (2011).

# 4 Results

This section will present the results for the DragonNet model under overlap violations. The first subsection will present the results using synthetic data. The next subsection will then show the results when using semi-synthetic datasets.

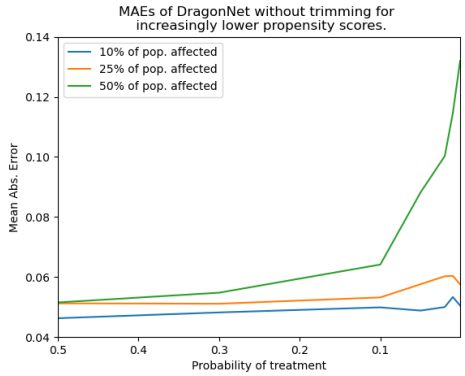
## 4.1 Synthetic results

First, the MAEs are obtained for increasingly lower propensity scores $P[T = 1|X_3 < w] \in [0.5, 0.3, 0.1, 0.05, 0.02, 0.01, 0.001]$, where $w \in [0.1, 0.25, 0.5]$ represents the fraction of the population affected by the worsening propensity scores. Worse propensity scores should make it harder to estimate the true ATE. Therefore, it is expected that for the extremely low propensity scores (, e.g., 1 % or less) the MAE should rise compared to the higher propensity scores. However, since the goal is to calculate the average treatment effect, it is expected that such bad propensity scores should not have a significant effect when only a small fraction (10%) of the population is affected, as the remaining 90% of the population could be sufficient to accurately calculate the ATE. If 25% or 50% of the population is affected, then there should be a significant increase in MAEs as a smaller part of the population remains with sufficient propensity scores to estimate the ATE.

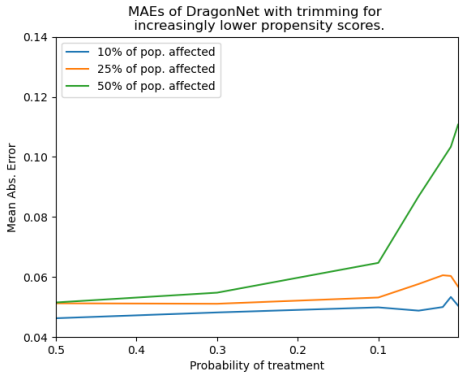
The results of the DragonNet with and without data trimming and the TARnet can be found in Figure 3. From the results of the DragonNet with and without trimming, it can be seen that they perform almost identically when 10% or 25% of the subpopulation is affected by extreme propensity scores. This is even the case for where the propensity scores drop below 0.05, which is when the data trimming might start to happen depending on the final estimated propensity scores for each point. The overall estimation performance of both models for extremely low propensity scores still seems to be relatively accurate when comparing it to the initial case where the propensity scores are still 0.5. Compared to the TARnet, both DragonNet versions seem outperform it, even when propensity scores are very low.

However, the performances of the three models start to differ somewhat significantly when a large portion (50%) of the population is affected. The MAE of the DragonNet without trimming starts to increase sharply for propensity scores lower than 0.1. When including the data trimming for the final ATE estimation, the MAE still seems to increase significantly, but the trimming actually manages to improve the MAE by a small margin for these smaller propensities. Both DragonNet versions, however, perform worse than the TARnet for these extreme propensity scores. The TARnet also has a consistently increasing MAE for these lower values, but the values are considerably lower than for the DragonNet models. The worse performance of the DragonNet is expected to be due to the targeted regularisation procedure, which modifies the outcome model and objective function by adding a term which contains estimated propensity scores in two denominators. Extreme propensity values could then lead to issues within the model's objective function and ATE estimation.

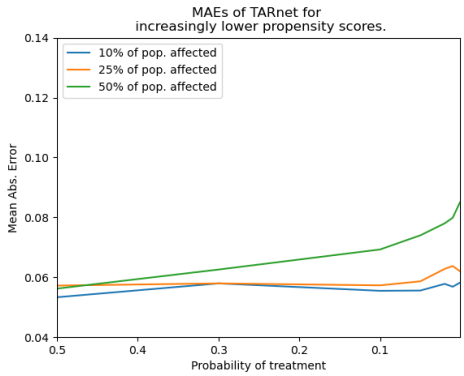
Next, the convergence results of the MAEs of the three models are obtained by increasing the sample size from 100 to 2000 while using different values for the propensity scores. As the convergence and convergence rate of the ATE estimator of the DragonNet depends on the convergence of the propensity score and outcome estimators, the expectation is that




(a) MAEs for DragonNet without trimming of datapoints with low estimated propensity.




(b) MAEs for DragonNet including trimming of datapoints with low estimated propensity.




(c) MAEs for default TARnet.

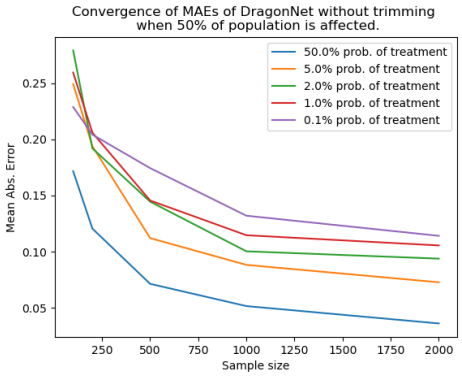
Figure 3: Graphs show effect of increasingly lower propensity scores on the MAE of the ATE estimators for the three different models using a sample size of 1000. The effects are visualised for when 10%, 25%, and 50% of the population is affected by the increasingly lower propensity scores, while the other portion has a 50% probability of treatment.

the convergence rate should become significantly worse when a large part of the population suffers from extremely low propensity scores. This should lead to a noticeable difference between the convergence rates when the propensity scores are 50% or very low (, e.g., 1% or less).

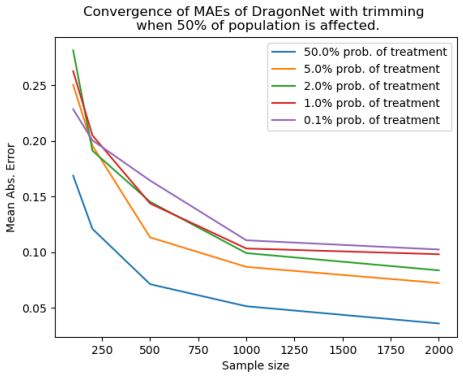
The results for when 50% of the population is affected can be found in Figure 4. It can be seen that the MAEs of both DragonNet versions still seem to drop significantly as the sample size increases. When there is no data trimming, however, the rate at which the MAE decreases seems to drop as the propensity score decreases compared to the case of a 50% propensity score. This is especially noticeable for when the propensity drops below 5%. If data trimming is performed, the MAEs seem to drop at a similar rate as 50% propensity again, although they still converge to a much higher MAE in the end, as is expected. The results for the TARnet, though, show that the model still has quickly decreasing MAEs in the case of low propensity scores, even faster than for the DragonNet with data trimming.

So, these results again show that the TARnet is able to handle worse propensity scores better than the DragonNet, likely due to the potential issues with the targeted regularisation in these extreme propensity cases.

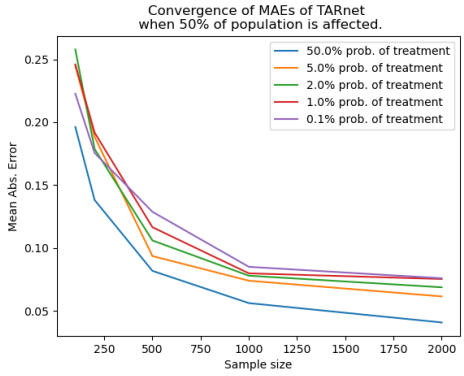
One final interesting observation which can be made from the convergence graphs of the DragonNet with and without trimming is that the MAEs of both models seem to converge to similar values, even though the trimming should lead to a biased estimate of the ATE. This result is due to the fact that the synthetic data has constant, homogeneous treatment effects. So, the ATE of any subpopulation is the same as the ATE of the overall population, which leads to the same results for the two models. If the treatment effects were not homogeneous, then the ATEs of the subpopulations would likely be different compared to the overall ATE and, thus, the results with and without trimming would also not converge to the same values anymore.




(a) MAEs for DragonNet without trimming of datapoints with low estimated propensity.




(b) MAEs for DragonNet including trimming of datapoints with low estimated propensity.




(c) MAEs for default TARnet.

Figure 4: Graphs show convergence of the MAE of the three models for different propensity scores when increasing the sample size. 50% of the population is affected by these decreasing propensity scores, while the rest of the population has a constant 50% probability of treatment.

Finally, the variability of the results over the 200 replications is obtained for the three

models. A lack of overlap in larger subpopulations makes it nearly impossible to estimate the treatment effect for that subpopulation, which should add additional uncertainty to the overall ATE estimate for the entire population. Therefore, when significant portions of the population suffer from extreme propensity scores, it is expected that the variance in the ATE estimations, and therefore MAEs, also increases.

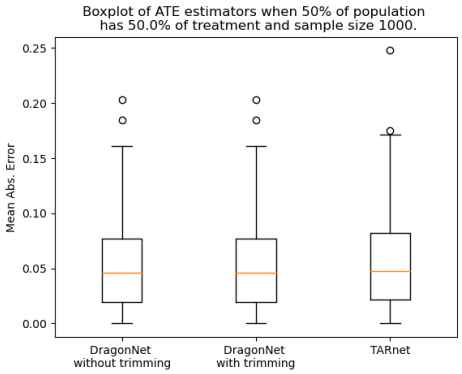
Figure 5 shows the box-plots of the results of the three models when 50% of the population is affected by bad propensity scores. When the subpopulation has a balanced 50% propensity score, the variance in the results between the three models seems to be almost identical. If the propensity score is reduced to 5%, some differences between the models appear. The DragonNet models now seem to have a larger spread in the results compared to before. The TARnet also has more variance in the results, but it is already less than for the DragonNet. Finally, in the most extreme case of 0.1% propensity, significant differences can be be observed between the models. The TARnet clearly has the least amount of variance in the results, roughly similar to the previous case of 5% propensity. The DragonNet without data trimming has almost twice the spread in its results compared to the TARnet. In case data trimming is added, it is slightly less, but still significantly more than TARnet. So, it seems that the TARnet can produce results with much less variance in the case of extreme propensity scores, which indicates that its results are not only more accurate than the DragonNet in these cases, but also more reliable.

## 4.2 Semi-synthetic results

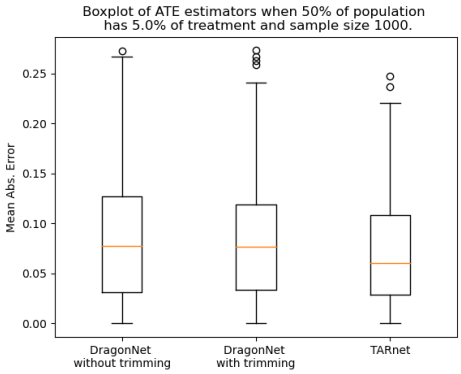
First, the results from running the three models on the increasingly imbalanced IHDP datasets are gathered. The imbalancing procedure affects the propensity scores of the entire population and not just a certain part of the population, as was the case in the synthetic data experiments (e.g., only 25% or 50% of the population). Therefore, it is expected that the performance of the DragonNet compared to the TARnet will be even worse on these datasets.

Figure 6 shows the results from using the IHDP datasets. Figure 6a shows that the DragonNet without trimming performs well at first, but as $q$ becomes larger than 0.5, the performance becomes significantly worse. The effect of trimming the bad propensity datapoints also seems to be quite significant, as the MAEs stay relatively close to those of the TARnet. As mentioned in the previous section, it is expected that this is due to the trimming of datapoints with low estimated propensity scores in the targeted regularisation term. The variation in the results, as shown in Figure 6b, also seem to reflect the performance differences. The variance of the results of the DragonNet without trimming is substantially higher compared to that of the DragonNet with trimming or the TARnet, and some large outliers can also be observed. It seems to be the case that if the entire population starts to suffer from poor overlap, the performance of the DragonNet quickly becomes unacceptable. Therefore, either trimming must be performed or another model, such as TARnet, should be used. The latter seems to be the best option, as even the DragonNet with trimming still performs worse than TARnet.

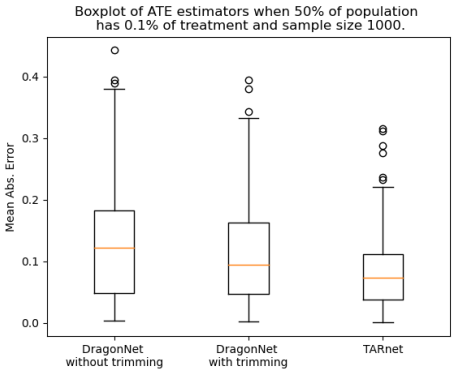
In order to support the claim that the poor performance of the DragonNet is largely due to the targeted regularisation term which uses the estimate propensity scores, Figure 7 shows the values of the term for all of the datapoints used with and without trimming for some random IHDP sample. From the results without trimming it can be observed that the term seems to be rather unstable when there is low overlap, as it attains very big positive or negative values. Trimming seems to completely remove all these large values from the data,




(a) Box-plots for when propensity is 50%.




(b) Box-plots for when propensity is 5%.



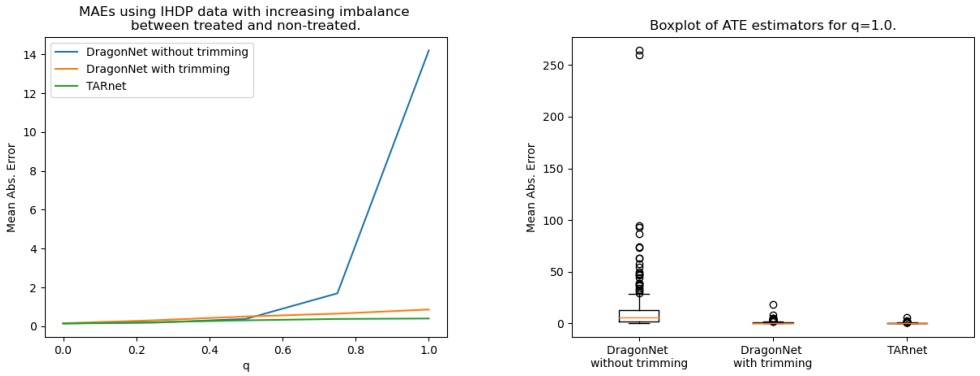

(c) Box-plots for when propensity is 0.1%.

Figure 5: The Box-plots show the variance in the MAEs of the three models for propensity scores of 50%, 5%, and 0.1%. 50% of the population is affected by the decreasing propensity scores. MAEs obtained over 200 replications with samples of size 1000.

which explains why it managed to obtain the significantly better results shown in Figure 6. So, from this experiment it does seem likely that the targeted regularisation procedure is one of the main causes of the poor DragonNet performance when there are (near) overlap violations.

# 5 Discussion

The results of the synthetic and semi-synthetic experiments showed that the DragonNet is not able to properly deal with datasets where significant portions of the population suffer from extreme propensity scores, especially compared to the TARnet. The data trimming based on estimated propensity scores when calculating the final ATE estimates seems to slightly improve the results in the simpler case where only a part of the population suffers from poor overlap. If the entire population starts to suffer, as was the case for the imbalanced IHDP datasets, the trimming starts to significantly improve the results. The TARnet, however, still produces much better and more reliable results in both situations. Therefore,




(a) MAEs of the three models with increasing amounts of imbalance.

(b) Box-plots of MAEs of the three models when imbalance parameter $q$ is set to 1.

Figure 6: Results of the DragonNet with and without trimming and the TARnet when ran on IHDP datasets with increasing amounts of imbalance. Overlap between treated and control populations is decreased as $q$ increases.

it seems to be better in most cases to use a different model than the DragonNet in case extreme propensity scores are observed for large portions of the population. However, in case only smaller portions suffer from bad propensity, for example about 25% of the population or less, the DragonNet still seems to be a suitable choice for estimating the ATE without requiring any trimming of data.

The poor performance of the DragonNet was expected to be due to the targeted regularisation procedure used by the model. Using a random IHDP sample to investigate the values of the specific term which uses estimated propensity scores, indicated that this may indeed be the cause of the poor performance, as very large positive and negative values were observed in the low overlap setting. Data trimming got rid of all these big outliers, and the DragonNet with trimming also managed to achieve significantly better results. Therefore, while targeted regularisation manages to increase performance in samples with (almost) no overlap issues (Shi et al., 2019), it also significantly hurts the performance in poor overlap situations.

One interesting observation from the results is the fact that the variance of the DragonNet results without data trimming is larger than when data trimming is applied in the extreme propensity cases. Petersen et al. (2010) suggested data trimming as a possible solution to deal with overlap violations, but noted that the trimming may lead to additional variance due to the resulting smaller sample size. In the experiments, however, it seems that the additional variability of the results due to extreme propensity scores heavily outweighs the extra variability resulting from a smaller sample, but with a lower degree of overlap violations. So, it seems that the DragonNet is more sensitive to extreme propensity scores than to smaller sample sizes.

However, even though data trimming seems to improve the result to some degree, it is important to note that applying this trimming changes the meaning of the final ATE estimate. Subpopulations with extreme propensity scores are now ignored in the final calculation. Therefore, the final ATE does not reflect the effect for the whole population anymore, but only for the portion with sufficient overlap. This might make the final result less inter-

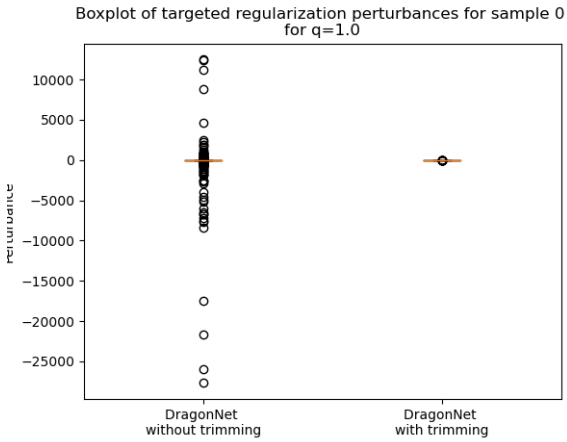



Figure 7: Box-plots of the values of the targeted regularisation terms which depend on the estimated propensity scores. Results shown for DragonNet with and without trimming on a random IHDP sample with imbalance parameter $q$ set to 1.

esting or useful in certain cases, especially when there are heterogeneous treatment effects in the population. Additionally, the performance of the DragonNet even with trimming might still not be good enough to warrant this change in the interpretation of the final results, especially considering that the results under overlap violations are still notably worse when compared to the TARnet.

While the results in this paper were obtained using one synthetic dataset and one semi-synthetic dataset (IHDP), the results can most likely still be generalised to many other scenarios. This is due to the fact that the poor performance of the DragonNet model under overlap violations seems to mainly be caused by the usage of estimated propensity scores in the targeted regularisation procedure. Even if other datasets have, for example, vastly different features or outcome models, they will still contain extreme propensity scores in the case of low overlap. Therefore, the DragonNet should still perform relatively badly in these other settings, as it will still obtain and use low estimates for the propensity scores.

# 6 Responsible Research

In order to help with the reproducibility of the results presented in this paper, the GitHub page with all functions used is publicly available [2]. The page also includes the 200 IHDP samples used, such that the results can be obtained again from the exact same datasets. Additionally, the functions used for generating synthetic datasets with different propensity scores and sizes are also provided. The settings used for obtaining the results were described in Section 3 and these can also be found in the code as the default settings used.

The ethical aspects of this work are related to groups of the population which may be underrepresented. Extreme propensity scores in data reflect the fact that some subpopulations with certain characteristics do not have enough overlap between treated and untreated individuals. It is therefore very difficult, or impossible, to accurately estimate the treatment

[2]Code can be found at: https://github.com/Marco-Murv/ResearchProject

effects for those subpopulations. This may lead to these underrepresented subpopulations either being ignored, in the case of data trimming, or possibly being assigned incorrect treatment effect estimations. Both of these cases are undesirable and can even potentially be dangerous if wrong treatment effects are assumed and then used. Therefore, understanding how the DragonNet performs under extreme propensity scores can be beneficial for avoiding these negative effects when estimating treatment effects by, for example, choosing other, more suitable methods in such cases.

# 7 Conclusion

The main question of this paper was about the performance of the DragonNet model under (near) overlap violations. It is clear that the model performs very poorly when large portions of the population suffer from low overlap, especially compared to other models, such as the TARnet. The poor performance seems to largely be due to estimated propensity scores being used in the targeted regularisation procedure. This also indicates that the results in this work can be generalised to many other low overlap settings, as those samples should always have low propensity scores in the data in those cases, no matter what other characteristics the samples have. Trimming the data by discarding the points with low estimated propensity scores seems to help, especially when the whole population suffers from overlap issues, but this leads to biased results which is not desirable, especially since the TARnet still performs better without any trimming. Overall, the best choice seems to be to use another model for estimating treatment effects, such as the TARnet, when it is suspected that large portions of the population suffer from low propensity.

Due to long computation times and limited available time overall, it was not possible to actually test the claimed generalisability of the results using other datasets or to test the performance against models which use completely different approaches to treatment effect estimation, instead of only using the TARnet which is somewhat similar to DragonNet. Therefore, it could be beneficial to extend the the experiments using a larger variety of datasets and models. This may also provide additional insights on the performance, and therefore usefulness, of other current state-of-the-art methods in the harder case of low overlap situations, as poor treatment effect estimations are undesirable and can potentially even be dangerous if used in practice. Besides simply testing more datasets and models, it could also be beneficial to investigate whether some metrics or procedures can be designed to indicate whether it is appropriate to use the DragonNet in a specific setting, as the model performs rather well under regular conditions when compared to other models.

# References


Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2016). Double/debiased machine learning for treatment and causal parameters.

Dorie, V. (2016). Npci: Non-parametrics for causal inference. https://github.com/vdorie/npci.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37.

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2009). *Causality*. Cambridge university press.

Petersen, M., Porter, K., Gruber, S., Wang, Y., and Laan, M. (2010). Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21:31–54.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Shadish, W., Cook, T., and Campbell, D. (2001). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.