Contrastive *Siamese Neural Networks* for conditional average treatment effects (CATE) estimation

Abstract

We propose a novel approach for estimating the *conditional average treatment effect* (CATE) by leveraging *Siamese Neural Networks* with a *contrastive learning* mechanism. The framework models treated and untreated pairs and learns a representation that captures the heterogeneity of treatment effects.

1 Problem Definition

Let Y denote a scalar *response* variable and Z denote a *binary treatment indicator* variable. Capital Roman letters denote random variables, while realized values appear in lower case, that is, y and z. Let x denote a length d vector of observed control variables. We will consider an observed sample of size n independent observations (Y_i, Z_i, x_i), for $i = 1, \ldots, n$. When Y or Z (respectively, y or z) are without a subscript, they denote length n column vectors; likewise, X will denote the $n \times d$ matrix of control variables.

We are interested in estimating various *treatment effects*. In particular, we are interested in *conditional average treatment effects* (CATE), the amount by which the response Y_i would differ between hypothetical worlds in which the treatment was set to $Z_i = 1$ versus $Z_i = 0$, averaged across subpopulations defined by attributes x. This kind of counterfactual estimate can be formalized in the *potential* outcomes framework by using $Y_i(0)$ and $Y_i(1)$ to denote the *observed* outcomes if treatment were set to zero or one, respectively. Under the *stable unit treatment value assumption* (SUTVA), we observe the potential outcome that corresponds to the realized treatment as

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

Throughout the paper we will assume that *strong ignorability* holds:

$$Yi(0), Yi(1) \perp\!\!\!\perp Zi|\mathbf{Xi}$$

and also that:

$$0 < Pr(Z_i = 1|x_i) < 1$$

for all $i = 1, \ldots, n$. The first condition assumes we have no unmeasured confounders, and the second condition (overlap) is necessary to estimate treatment effects everywhere in covariate space. Provided that these conditions hold, it follows that $E(Y_i(z)|xi) = E(Y_i|x_i, Z_i = z)$ so our estimate is:

$$\tau(x_i) = E(Y_i|x_i, Z_i = 1) - E(Y_i|x_i, Z_i = 0)$$

2 Siamese Neural Network framework

To estimate the CATE $\tau(\boldsymbol{x}_i)$, we leverage siamese neural networks (SNNs), which are designed to learn a shared representation $\phi(\boldsymbol{x_i})$ for pairs of individuals. This shared representation captures the most relevant features of the covariates $\boldsymbol{x}_i$ while suppressing irrelevant or confounding information, enabling meaningful comparisons between treated and untreated groups.

The SNN processes each individual's covariates $\boldsymbol{x}_i$ through the same network (i.e., shared weights), producing *embeddings* $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{x}_j)$ for any pair of individuals $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The shared representation $\phi(\boldsymbol{x})$ ensures that the embedding space is consistent and comparable for all individuals, facilitating the computation of distances between pairs:

$$d_{ij} = \|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|_2, \tag{1}$$

where d_{ij} measures the similarity or dissimilarity between the embeddings.

To guide the learning process, pairs of individuals are classified as:

- *positive pairs:* $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ such that $Z_i = Z_j$, i.e., both individuals are either treated ($Z_i = Z_j = 1$) or untreated ($Z_i = Z_j = 0$); the network is trained to minimize the distance d_{ij}, encouraging similar embeddings.

- *negative pairs:* $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ such that $Z_i \neq Z_j$, i.e., one individual is treated ($Z_i = 1$) and the other is untreated ($Z_j = 0$); the network is trained to maximize the distance d_{ij}, up to a predefined margin m, ensuring that embeddings of treated and untreated individuals are distinguishable.

The shared representation $\phi(\boldsymbol{x})$ learned serves several critical purposes: *(i) feature extraction*, i.e., the covariates most relevant for the treatment effect

estimation, while suppressing irrelevant or noisy information, *(ii) pairwise comparisons*, i.e., by encoding individuals into a consistent embedding space, the network enables meaningful comparisons between treated and untreated groups, *(iii) improved CATE estimation*, i.e., the learned representation simplifies the modeling of potential outcomes $Y_i(1)$ and $Y_i(0)$ by emphasizing the information needed to distinguish between treated and untreated individuals.

The goal of the *contrastive loss function* is to optimize embeddings to:

- *minimize* the distance d_{ij} for positive pairs.
- *maximize* the distance d_{ij} for negative pairs, up to a *margin* m, where m is a hyperparameter that defines the minimum acceptable distance between embeddings of negative pairs; it plays a critical role in shaping the embedding space learned by SNN.

The *loss function* can be formalized as follows:

$$\mathcal{L}_{\text{contrastive}} = \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)} \left[\mathbb{I}_{\{Z_i = Z_j\}} \cdot d_{ij}^2 + \mathbb{I}_{\{Z_i \neq Z_j\}} \cdot \max(0, m - d_{ij})^2 \right]. \tag{2}$$

where the *indicator function* $\mathbb{I}_{\{Z_i = Z_j\}}$ is defined as:

$$\mathbb{I}_{\{Z_i = Z_j\}} = \begin{cases} 1, & \text{if } Z_i = Z_j, \\ 0, & \text{if } Z_i \neq Z_j. \end{cases} \tag{3}$$

The notation $\mathbb{I}_{\{Z_i \neq Z_j\}}$ is defined in a similar and specular way.

The *expected potential outcomes* for individual i, given its covariates $\boldsymbol{x}_i$, are defined $\mathbb{E}[Y_i(1) \mid \boldsymbol{x}_i] = g(\phi(\boldsymbol{x}_i), Z_i = 1)$ and $\mathbb{E}[Y_i(0) \mid \boldsymbol{x}_i] = g(\phi(\boldsymbol{x}_i), Z_i = 0)$, where $\mathbb{E}[Y_i(1) \mid \boldsymbol{x}_i]$ is the expected outcome for individual i if treated, and $\mathbb{E}[Y_i(0) \mid \boldsymbol{x}_i]$ is the expected outcome if untreated.

The CATE is estimated as the difference of expected potential outcomes:

$$t(\boldsymbol{x}_i) = \mathbb{E}[Y_i(1) \mid \boldsymbol{x}_i] - \mathbb{E}[Y_i(0) \mid \boldsymbol{x}_i]. \tag{4}$$

Observe that the SNN learns a shared representation $\phi(\boldsymbol{x}_i)$ for each individual i, where $\phi : \mathbb{R}^d \to \mathbb{R}^k$ maps the covariates $\boldsymbol{x}_i$ to a lower-dimensional embedding space. Such embeddings $\phi(\boldsymbol{x}_i)$ are optimized to: *(i)* bring embeddings of individuals in the same treatment group ($Z_i = Z_j$) closer together, *(ii)* push embeddings of individuals in different treatment groups ($Z_i \neq Z_j$) farther apart, up to a margin m. After training, $\phi(\boldsymbol{x}_i)$ encodes the most relevant features of the covariates $\boldsymbol{x}_i$ for estimating treatment effects.

To estimate the expected potential outcomes for each individual, we can use a *regression* model $g(\cdot)$ that takes the embedding $\phi(\boldsymbol{x}_i)$ and the treatment

indicator Z_i as inputs. The regression function $g(\phi(\boldsymbol{x}_i), Z_i)$ approximates the conditional expected value of the outcome Y_i:

$$\mathbb{E}[Y_i(1) \mid \boldsymbol{x}_i] \approx g(\phi(\boldsymbol{x}_i), Z_i = 1), \tag{5}$$

$$\mathbb{E}[Y_i(0) \mid \boldsymbol{x}_i] \approx g(\phi(\boldsymbol{x}_i), Z_i = 0). \tag{6}$$

The model $g(\cdot)$ can be implemented using various regression techniques, such as, a *Linear regression*, *Fully connected neural networks* (MLPs), *Tree-based methods*, such as random forests, and is trained using the observed outcomes Y_i and the treatment assignments Z_i, with the goal of minimizing the prediction error on the observed data:

$$\mathcal{L}_{\text{regression}} = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - g(\phi(\boldsymbol{x}_i), Z_i)\big)^2, \tag{7}$$

where Y_i is the observed outcome for individual i, and $g(\phi(\boldsymbol{x}_i), Z_i)$ is the predicted outcome for individual i under the observed treatment Z_i.

Once $g(\cdot)$ is trained, we use it to predict the potential outcomes for i as:

$$\hat{Y}_i(1) = g(\phi(\boldsymbol{x}_i), Z_i = 1), \tag{8}$$

$$\hat{Y}_i(0) = g(\phi(\boldsymbol{x}_i), Z_i = 0). \tag{9}$$

where $\hat{Y}_i(1)$ is the predicted outcome for individual i if treated, and $\hat{Y}_i(0)$ is the predicted outcome for individual i if untreated. Thus, the CATE is estimated as the difference between the predicted potential outcomes:

$$\hat{t}(\boldsymbol{x}_i) = \hat{Y}_i(1) - \hat{Y}_i(0). \tag{10}$$

The use of a regression model $g(\cdot)$ in combination with the SNN offers several advantages: *(i) flexibility*, i.e., the regression model can be chosen based on the complexity of the relationship between covariates, treatment, and outcomes, *(ii) interpretability*, i.e., by separating the embedding ($\phi(\boldsymbol{x}_i)$) from the outcome prediction, the method provides interpretable representations of the covariates, *(iii) efficiency*, i.e., the embeddings $\phi(\boldsymbol{x}_i)$ reduce the dimensionality of the input, simplifying the regression model and improving computational efficiency, and finally *(iv) generalizability*, i.e., the SNN ensures that the embeddings generalize well across treated and untreated groups, enabling robust CATE estimation.

Here we provide a proof-of-concept scheme of how integrating the contrastive embedding mechanism and regression-based expected outcome estimation into a unified framework can be used for CATE estimation.

Proposition 2.1. *Let $\phi(\boldsymbol{x}_i)$ be the embedding learned using a contrastive loss $\mathcal{L}_{contrastive}$, which ensures that:*

- *Positive pairs (same treatment group, $Z_i = Z_j$) are mapped closer together in the embedding space: $\|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|_2$ is small.*
- *Negative pairs (different treatment groups, $Z_i \neq Z_j$) are separated by at least a margin m: $\|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|_2 \geq m$.*

Let $g(\phi(\boldsymbol{x}_i), Z_i)$ be a regression model that is trained to minimize:

$$\mathcal{L}_{regression} = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - g(\phi(\boldsymbol{x}_i), Z_i)\big)^2. \tag{11}$$

If $g(\cdot)$ is sufficiently expressive, it approximates the true conditional expectation:

$$g(\phi(\boldsymbol{x}_i), Z_i) \to \mathbb{E}[Y_i \mid \boldsymbol{x}_i, Z_i] \quad \text{as } n \to \infty. \tag{12}$$

Proof. The contrastive loss ensures that the learned embeddings $\phi(\boldsymbol{x}_i)$ encode treatment-relevant information from $\boldsymbol{x}_i$:

- $Z_i = Z_j$: for individuals with similar covariates in the same treatment group, the contrastive loss minimizes $\|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|_2$, ensuring that embeddings of similar treated/untreated individuals are close together.
- $Z_i \neq Z_j$: for individuals with similar covariates in different treatment groups, the loss enforces $\|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|_2 \geq m$, ensuring separation between treated/untreated groups in the embedding space.

This implies that $\phi(\boldsymbol{x}_i)$ preserves treatment-relevant covariate information while suppressing irrelevant variability, making $\phi(\boldsymbol{x}_i)$ a sufficient representation of $\boldsymbol{x}_i$ for treatment effect estimation. Since $\phi(\boldsymbol{x}_i)$ preserves the treatment-relevant information from $\boldsymbol{x}_i$, we have:

$$\mathbb{E}[Y_i \mid \phi(\boldsymbol{x}_i), Z_i] = \mathbb{E}[Y_i \mid \boldsymbol{x}_i, Z_i].$$

This follows from the fact that $\phi(\boldsymbol{x}_i)$ is a deterministic function of $\boldsymbol{x}_i$, and the embedding separates treated and untreated groups effectively.

Observe that the regression model $g(\phi(\boldsymbol{x}_i), Z_i)$ is trained to minimize:

$$\mathcal{L}_{\text{regression}} = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - g(\phi(\boldsymbol{x}_i), Z_i)\big)^2$$

so, by the *Universal Approximation Theorem* for regression models (e.g., neural networks), $g(\cdot)$ can approximate any measurable function of its inputs, and in our case it can approximate the true conditional expectation:

$$g(\phi(\boldsymbol{x}_i), Z_i) \to \mathbb{E}[Y_i \mid \phi(\boldsymbol{x}_i), Z_i] \quad \text{as } n \to \infty. \tag{13}$$

As a consequence, we have the following result:

$$g(\phi(\boldsymbol{x}_i), Z_i) \to \mathbb{E}[Y_i \mid \phi(\boldsymbol{x}_i), Z_i] = \mathbb{E}[Y_i \mid \boldsymbol{x}_i, Z_i].$$

□

Proposition 2.2. *Let $\phi(\boldsymbol{x}_i)$ be a representation learned using the loss $\mathcal{L}_{contrastive}$, and let $g(\phi(\boldsymbol{x}_i), Z_i)$ be a regression function trained to minimize:*

$$\mathcal{L}_{regression} = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - g(\phi(\boldsymbol{x}_i), Z_i)\big)^2. \tag{14}$$

Under the following assumptions:

1. *The treatment assignment is independent of potential outcomes given covariates, i.e., $Y_i(1), Y_i(0) \perp Z_i \mid \boldsymbol{x}_i$,*
2. *The probability of receiving treatment is strictly between 0 and 1 for all individuals, i.e., $0 < \Pr(Z_i = 1 \mid \boldsymbol{x}_i) < 1$,*
3. *The regression model $g(\cdot)$ is sufficiently flexible to approximate the true conditional expectations, i.e., $g(\phi(\boldsymbol{x}_i), Z_i) \approx \mathbb{E}[Y_i \mid \boldsymbol{x}_i, Z_i]$.*

The estimated Conditional Average Treatment Effect (CATE):

$$\hat{t}(\boldsymbol{x}_i) = g(\phi(\boldsymbol{x}_i), Z_i = 1) - g(\phi(\boldsymbol{x}_i), Z_i = 0), \tag{15}$$

converges to the true CATE $t(\boldsymbol{x}_i) = \mathbb{E}[Y_i(1) \mid \boldsymbol{x}_i] - \mathbb{E}[Y_i(0) \mid \boldsymbol{x}_i]$, as the sample size $n \to \infty$.

Proof. We recall that the contrastive loss is defined as:

$$\mathcal{L}_{\text{contrastive}} = \sum_{(i,j)} \left[\mathbb{I}_{\{Z_i = Z_j\}} d_{ij}^2 + \mathbb{I}_{\{Z_i \neq Z_j\}} \max(0, m - d_{ij})^2\right]$$

where $d_{ij} = \|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|_2$ is the distance between embeddings. Training the SNN minimizes this loss, ensuring that embeddings of treated and untreated individuals ($Z_i \neq Z_j$) are pushed apart by at least the margin m, while embeddings of individuals in the same treatment group ($Z_i = Z_j$) are

pulled closer together. By separating treated and untreated groups in the embedding space, $\phi(\boldsymbol{x}_i)$ preserves covariate information relevant to treatment effects while suppressing irrelevant variability.

The regression model $g(.)$ is trained on the observed Y_i to minimize:

$$\mathcal{L}_{\text{regression}} = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - g(\phi(\boldsymbol{x}_i), Z_i)\big)^2. \tag{16}$$

Since $g(\cdot)$ is sufficiently expressive (by Assumption 3), by Proposition 2.1, it learns to approximate the true conditional expectations:

$$g(\phi(\boldsymbol{x}_i), Z_i) \to \mathbb{E}[Y_i \mid \boldsymbol{x}_i, Z_i] \quad \text{as } n \to \infty. \tag{17}$$

Using $g(\cdot)$, we estimate the expected potential outcomes:

$$\mathbb{E}[Y_i(1) \mid \boldsymbol{x}_i] \approx g(\phi(\boldsymbol{x}_i), Z_i = 1)$$

$$\mathbb{E}[Y_i(0) \mid \boldsymbol{x}_i] \approx g(\phi(\boldsymbol{x}_i), Z_i = 0)$$

The CATE is defined as $t(\boldsymbol{x}_i) = \mathbb{E}[Y_i(1) \mid \boldsymbol{x}_i] - \mathbb{E}[Y_i(0) \mid \boldsymbol{x}_i]$, so by substituting the regression estimates we obtain $\hat{t}(\boldsymbol{x}_i) = g(\phi(\boldsymbol{x}_i), Z_i = 1) - g(\phi(\boldsymbol{x}_i), Z_i = 0)$. As $n \to \infty$, the consistency of the regression model ensures:

$$\hat{t}(\boldsymbol{x}_i) \to t(\boldsymbol{x}_i) \tag{18}$$

□