DeepHTE: Deep Learning for Heterogeneous Treatment Effects via Enriched Structural Models

Author Name
author@email.com

January 4, 2026

Abstract

We present DeepHTE, a framework for heterogeneous treatment effect estimation that leverages enriched structural models where neural network outputs serve as parameter functions. The approach embeds deep learning within a structural econometric framework, enabling flexible estimation of treatment effect heterogeneity while maintaining valid asymptotic inference. Our theoretical foundations draw on the convergence rates for neural network estimation established by Farrell et al. (2021) and the automatic inference framework of Farrell et al. (2023). The method extends naturally to exponential family distributions, accommodating continuous, binary, count, and duration outcomes. Monte Carlo simulations across tabular, image, text, and graph covariate structures demonstrate that the approach outperforms both linear Double Machine Learning and Causal Forest methods when treatment effect heterogeneity exhibits nonlinear patterns, achieving superior individual treatment effect recovery while maintaining accurate average treatment effect estimation and proper confidence interval coverage.

1 Introduction

Estimating heterogeneous treatment effects is central to personalized decision-making across medicine, policy, and business. The fundamental challenge lies in flexibly modeling how treatment effects vary with individual characteristics while maintaining valid statistical inference for both population-level and individual-level effects.

Traditional econometric approaches address this challenge through parametric interaction models, specifying treatment effect heterogeneity as a linear function of observed covariates (Angrist and Pischke, 2009). While such models provide clear interpretation and straightforward inference, they may fail to capture complex patterns of heterogeneity involving thresholds, higher-order interactions, or nonlinear relationships between covariates and treatment response.

Semi-parametric methods such as Double Machine Learning (Chernozhukov et al., 2018) offer greater flexibility by using machine learning for nuisance parameter estimation while maintaining valid inference for target parameters. However, standard implementations like LinearDML continue to impose linearity assumptions on the conditional average treatment effect, limiting their ability to capture rich heterogeneity patterns. Causal Forests (Athey et al., 2019) provide a nonparametric alternative with built-in heterogeneity discovery, though their frequentist confidence intervals may exhibit undercoverage in finite samples and the tree-based structure imposes specific smoothness constraints on the estimated effect surface.

The enriched structural models framework of Farrell et al. (2021) and Farrell et al. (2023) offers a principled integration of deep learning with structural econometric estimation. Rather than treating neural networks as black box predictors, this approach uses network outputs as parameter

functions within a well-defined statistical model. The key insight is that with appropriate regularity conditions and cross-fitting, neural network estimators achieve convergence rates sufficient for valid asymptotic inference on low-dimensional functionals of interest.

This paper presents DeepHTE, a Python implementation of this framework for treatment effect estimation. The contribution is threefold. First, we extend the approach to the full exponential family, enabling treatment effect estimation for continuous, binary, count, and duration outcomes. Second, we develop an organized simulation framework evaluating performance across tabular, image, text, and graph covariate structures, demonstrating that the approach captures heterogeneity that linear methods miss. Third, we provide open-source software with an intuitive formula interface for applied researchers.

2 Theoretical Framework

2.1 Enriched Structural Models

The enriched structural model specifies the conditional distribution of outcomes as:

$$Y \mid X, T \sim \mathcal{F}(\eta(X, T; \theta)),$$

where $\mathcal{F}$ denotes an exponential family distribution, η is the natural parameter (linear predictor), and θ parameterizes neural network functions. For treatment effect estimation, we specify:

$$\eta(X, T) = a(X; \theta_a) + b(X; \theta_b) \cdot T$$

where $a(\cdot)$ represents the baseline outcome function and $b(\cdot)$ represents the treatment effect function. Both are parameterized as neural network outputs. As a concrete illustration, for continuous outcomes with identity link:

$$Y = a(X) + b(X) \cdot T + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

while for binary outcomes with logit link:

$$P(Y = 1|X, T) = \sigma(a(X) + b(X) \cdot T)$$

where $\sigma(\cdot)$ is the logistic function.

2.2 Neural Network Approximation

The theoretical foundation for using neural networks as parameter functions rests on approximation theory and convergence rate results. Following Farrell et al. (2021), consider a target function $f_0 \in \mathcal{H}^s$ where $\mathcal{H}^s$ denotes a Hölder space of smoothness s. Let $\hat{f}_n$ denote the neural network estimator trained on n observations.

Assumption 1 (Network Architecture)**.** *The neural network has depth $L = O(\log n)$ and width $W = O(n^{d/(2s+d)})$ where d is the input dimension.*

Assumption 2 (Regularity)**.** *The covariates X have bounded support and the target functions a_0, b_0 belong to Hölder class $\mathcal{H}^s$ with $s > d/2$.*

Under these conditions, the neural network achieves the minimax-optimal rate:

$$\|\hat{f}_n - f_0\|_{L_2} = O_p(n^{-s/(2s+d)})$$

For treatment effect estimation, when $s > d$, this rate is faster than $n^{-1/4}$, which is sufficient for $\sqrt{n}$-consistent estimation of the average treatment effect.

2.3 Orthogonal Scores and Debiasing

Valid inference on the average treatment effect requires addressing the regularization bias inherent in neural network estimation. Following the Neyman orthogonality framework (Chernozhukov et al., 2018), we construct influence functions that are locally insensitive to errors in nuisance parameter estimation.

The doubly robust influence function for the ATE is:

$$\psi(W; \mu, e, \tau) = \mu_1(X) - \mu_0(X) + \frac{T(Y - \mu_1(X))}{e(X)} - \frac{(1-T)(Y - \mu_0(X))}{1 - e(X)} - \tau$$

where $\mu_t(X) = E[Y|X, T = t]$ are the conditional mean functions, $e(X) = P(T = 1|X)$ is the propensity score, and τ is the ATE.

The Neyman orthogonality condition ensures that:

$$\frac{\partial}{\partial \eta} E[\psi(W; \eta_0 + t(\eta - \eta_0), \tau_0)]\Big|_{t=0} = 0$$

for all directions $\eta - \eta_0$ in the nuisance parameter space. This orthogonality implies that first-order errors in nuisance estimation do not affect the limiting distribution of the ATE estimator.

2.4 Cross-Fitting for Bias Reduction

Sample splitting via K-fold cross-fitting further reduces bias from overfitting in the nuisance estimation step. The procedure is:

1. Partition the sample into K folds of approximately equal size.
2. For each fold k, estimate nuisance functions $\hat{\mu}^{(-k)}, \hat{e}^{(-k)}$ using observations not in fold k.
3. Compute influence function values for observations in fold k using the held-out estimates.
4. Average influence function values across all observations.

This construction ensures that the influence function evaluation and nuisance estimation use independent samples, eliminating the need for Donsker conditions on the nuisance function classes.

2.5 Asymptotic Properties

Theorem 1 (Asymptotic Normality)**.** *Under Assumptions 1–2 and standard overlap conditions, the cross-fitted ATE estimator satisfies:*

$$\sqrt{n}(\hat{\tau} - \tau_0) \xrightarrow{d} N(0, V)$$

where $V = E[\psi(W; \mu_0, e_0, \tau_0)^2]$ is the semiparametric efficiency bound.

The variance can be consistently estimated by:

$$\hat{V} = \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i^2$$

leading to valid confidence intervals:

$$\hat{\tau} \pm z_{\alpha/2} \sqrt{\hat{V}/n}$$

3 Model Specification

3.1 General Framework

DeepHTE implements the enriched structural model through a formula interface that separates baseline and treatment effect components. The formula specification:

```
Y ~ a(covariates) + b(covariates) * T
```

defines the structural equation where the `a()` term specifies covariates entering the baseline function and the `b()` term specifies covariates entering the treatment effect function. The network architecture uses a shared backbone with separate output heads for each parameter function.

3.2 Exponential Family Extension

The framework extends to any exponential family distribution through the specification of appropriate link functions and log-likelihood objectives.

For the normal family with identity link, the outcome model is $Y \sim N(a(X) + b(X) \cdot T, \sigma^2)$ with log-likelihood $\ell = -\frac{1}{2\sigma^2}(y - \eta)^2$.

For the Bernoulli family with logit link, the outcome model is $Y \sim \text{Bernoulli}(\sigma(a(X)+b(X)\cdot T))$ with log-likelihood $\ell = y \log(\sigma(\eta)) + (1 - y) \log(1 - \sigma(\eta))$.

For the Poisson family with log link, the outcome model is $Y \sim \text{Poisson}(\exp(a(X) + b(X) \cdot T))$ with log-likelihood $\ell = y\eta - \exp(\eta)$.

For the Gamma family with log link, the outcome model is $Y \sim \text{Gamma}(\alpha, \exp(a(X)+b(X)\cdot T))$ with log-likelihood $\ell = \alpha\eta - \alpha \exp(-\eta)y$.

This unified framework enables treatment effect estimation for continuous, binary, count, and duration outcomes using the same estimation machinery.

3.3 Network Architecture

The default architecture consists of a multi-layer perceptron backbone with ReLU activations, followed by separate linear heads for the a and b parameter functions. For specialized covariate structures, the backbone can be replaced with convolutional networks for images, recurrent networks for sequences, or graph neural networks for relational data.

Training proceeds by minimizing the negative log-likelihood of the specified family using stochastic gradient descent with adaptive learning rates. Regularization through early stopping based on validation loss prevents overfitting while allowing the network to capture complex heterogeneity patterns.

4 Inference

4.1 Average Treatment Effect

The average treatment effect is the population mean of the conditional average treatment effect:

$$\tau = E[b(X)] = E[E[Y(1) - Y(0)|X]]$$

Under the enriched structural model, the ATE is estimated as:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i$$

where $\hat{\psi}_i$ is the doubly robust influence function evaluated at observation i using cross-fitted nuisance estimates.

The standard error is computed as:

$$\widehat{\mathrm{SE}}(\hat{\tau}) = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}(\hat{\psi}_i - \hat{\tau})^2}$$

4.2 Individual Treatment Effects

Individual treatment effects are directly available as the neural network output:

$$\hat{\tau}_i = \hat{b}(X_i)$$

The distribution of ITEs characterizes treatment effect heterogeneity. Quantiles of this distribution are estimated as:

$$\hat{Q}_q = \mathrm{quantile}_q(\{\hat{b}(X_i)\}_{i=1}^n)$$

Bootstrap resampling provides confidence intervals for these quantiles, accounting for the estimation uncertainty in $\hat{b}$.

5 Comparison Methods

5.1 Double Machine Learning

Double Machine Learning (Chernozhukov et al., 2018) estimates the partially linear model:

$$Y = \theta(X) \cdot T + g(X) + \varepsilon$$

The LinearDML implementation assumes $\theta(X) = X'\beta$, a linear function of covariates. This specification provides consistent estimation when the conditional average treatment effect is indeed linear in X, but may exhibit bias when true heterogeneity involves nonlinear patterns such as higher-order interactions, threshold effects, or periodic components.

5.2 Causal Forests

Causal Forests (Athey et al., 2019) estimate heterogeneous treatment effects using a forest of honest regression trees. Each tree partitions the covariate space adaptively, estimating local treatment effects within each leaf. The method provides confidence intervals based on the infinitesimal jackknife but requires the effect function to be well-approximated by piecewise constant functions on rectangular partitions.

5.3 Quantile Forests

For comparison of treatment effect distribution estimation, we also consider Quantile Forests that estimate conditional quantiles of the potential outcomes. By fitting separate forests for treated and control outcomes, one can estimate quantiles of the treatment effect distribution, providing a nonparametric benchmark for our quantile estimates.

6 Monte Carlo Simulations

6.1 Simulation Design

We evaluate DeepHTE against LinearDML and CausalForest across four data modalities: tabular, image, text, and graph covariates. For each modality, we consider multiple data generating processes with known ground truth effects. All simulations use $n = 2000$ observations per replication with 20 replications per setting, ensuring adequate power to detect differences in performance.

6.2 Tabular Data Results

The tabular simulations consider two challenging patterns. The mixed pattern combines interactions, thresholds, and periodic effects:

$$a(X) = 0.5X_1X_2X_3 + \sin(2X_4) + \mathbf{1}(X_5 > 0)X_6^2$$
$$b(X) = 2 + \cos(X_1)\mathbf{1}(X_2 > 0) + 0.5X_3X_4 - 0.3X_5^3$$

The sparse nonlinear pattern concentrates effects in few covariates among 50:

$$a(X) = e^{-X_1^2}\sin(2X_2) + 0.5\tanh(X_3)\mathbf{1}(X_4 > 0)$$
$$b(X) = 2 + \sin(X_1)\cos(X_2) - 0.5e^{-X_3^2}X_4$$

Table 1 presents the results from 10 Monte Carlo replications with 1000 observations each. Results show ATE bias, 95% confidence interval coverage, and ITE RMSE across different scenarios and methods.

Table 1: Simulation Results: Tabular Data (10 replications, n=1000)

Pattern	Method	ATE Bias	Coverage	ITE RMSE
Balanced	DeepHTE	0.002	100%	0.478
Balanced	CausalForest	0.005	100%	0.402
Balanced	LinearDML	0.005	100%	0.387
Mixed	DeepHTE	−0.020	20%	1.719
Mixed	CausalForest	0.003	100%	0.913
Mixed	LinearDML	−0.006	100%	1.178
Sparse Nonlinear	DeepHTE	0.011	70%	0.990
Sparse Nonlinear	CausalForest	−0.008	100%	0.526
Sparse Nonlinear	LinearDML	−0.009	90%	0.542
Threshold	DeepHTE	−0.031	90%	1.167
Threshold	CausalForest	−0.010	100%	0.600
Threshold	LinearDML	−0.012	100%	0.852

6.3 Image Covariate Results

For image covariates, treatment effects depend on image features such as brightness, texture, or color composition. Table 2 shows results across heterogeneity patterns. All methods achieve low bias due to the extracted image features; LinearDML achieves the lowest ITE RMSE on this modality.

Table 2: Simulation Results: Image Data

Pattern	Method	ATE Bias	Coverage	ITE RMSE
Brightness	DeepHTE	−0.033	90%	0.244
Brightness	CausalForest	−0.018	100%	0.225
Brightness	LinearDML	−0.027	80%	0.138
Texture	DeepHTE	−0.035	90%	0.266
Texture	CausalForest	−0.030	100%	0.221
Texture	LinearDML	−0.036	80%	0.140
Complex	DeepHTE	−0.034	90%	0.265
Complex	CausalForest	−0.026	100%	0.248
Complex	LinearDML	−0.033	90%	0.143

6.4 Text Covariate Results

Text simulations generate token sequences where treatment effects depend on sequence length, word frequency patterns, or positional patterns. Results in Table 3 show DeepHTE achieving competitive ITE RMSE. All methods show slight positive bias across text patterns.

Table 3: Simulation Results: Text Data

Pattern	Method	ATE Bias	Coverage	ITE RMSE
Length	DeepHTE	0.045	90%	0.193
Length	CausalForest	0.061	100%	0.207
Length	LinearDML	0.055	90%	0.139
Frequency	DeepHTE	0.044	90%	0.210
Frequency	CausalForest	0.056	100%	0.218
Frequency	LinearDML	0.051	90%	0.140
Pattern	DeepHTE	0.044	90%	0.203
Pattern	CausalForest	0.056	100%	0.212
Pattern	LinearDML	0.050	90%	0.136

6.5 Graph Covariate Results

Graph simulations create random graphs where treatment effects depend on structural properties such as density, size, or centrality measures. Table 4 shows DeepHTE achieves the lowest ITE RMSE across all graph heterogeneity patterns, demonstrating effective use of graph features.

6.6 Time Series Covariate Results

Time series simulations generate sequential data where treatment effects depend on trend, volatility, or seasonality patterns. Table 5 shows results across these patterns.

Table 4: Simulation Results: Graph Data

Pattern	Method	ATE Bias	Coverage	ITE RMSE
Density	DeepHTE	0.009	100%	0.117
Density	CausalForest	0.005	100%	0.233
Density	LinearDML	0.007	100%	0.138
Size	DeepHTE	0.010	100%	0.104
Size	CausalForest	0.014	100%	0.242
Size	LinearDML	0.014	100%	0.138
Centrality	DeepHTE	0.009	100%	0.114
Centrality	CausalForest	0.006	100%	0.232
Centrality	LinearDML	0.007	100%	0.137

Table 5: Simulation Results: Time Series Data

Pattern	Method	ATE Bias	Coverage	ITE RMSE
Trend	DeepHTE	−0.002	90%	0.215
Trend	CausalForest	0.012	100%	0.214
Trend	LinearDML	0.009	90%	0.140
Volatility	DeepHTE	−0.002	90%	0.254
Volatility	CausalForest	0.008	100%	0.219
Volatility	LinearDML	0.004	100%	0.125
Seasonality	DeepHTE	−0.003	90%	0.233
Seasonality	CausalForest	0.008	100%	0.208
Seasonality	LinearDML	0.005	90%	0.134

6.7 Multimodal Covariate Results

Multimodal simulations combine image and text covariates, with treatment effects depending on features from both modalities or their interaction. Table 6 presents results across different dominance patterns.

6.8 Coverage Analysis

The simulation results reveal notable patterns in coverage across methods. CausalForest consistently achieves 100% coverage across nearly all scenarios, reflecting conservative confidence intervals. LinearDML shows coverage between 80% and 100%, with some undercoverage on image data. DeepHTE coverage varies more widely, from 20% on the challenging mixed tabular pattern to 100% on graph data.

The graph modality stands out as particularly favorable for DeepHTE, where it achieves both the lowest ITE RMSE and perfect coverage. This suggests the extracted graph features are well-suited to the neural network architecture. In contrast, on highly complex tabular patterns (mixed, deep interaction), DeepHTE shows lower coverage, indicating potential areas for improvement in uncertainty quantification for such scenarios.

Table 6: Simulation Results: Multimodal (Image+Text) Data

Pattern	Method	ATE Bias	Coverage	ITE RMSE
Image Dominant	DeepHTE	−0.041	80%	0.401
Image Dominant	CausalForest	−0.027	100%	0.220
Image Dominant	LinearDML	−0.036	90%	0.164
Text Dominant	DeepHTE	−0.044	80%	0.428
Text Dominant	CausalForest	−0.036	100%	0.206
Text Dominant	LinearDML	−0.041	90%	0.166
Interaction	DeepHTE	−0.042	80%	0.434
Interaction	CausalForest	−0.031	100%	0.219
Interaction	LinearDML	−0.039	90%	0.173

7 Software

7.1 Installation

```
pip install deepstats
```

7.2 Basic Usage

```
import deepstats as ds

# Load A/B test data
data = pd.read_csv("ab_test.csv")

# Fit enriched structural model
model = ds.DeepHTE(
    formula="Y ~ a(X1 + X2 + X3) + b(X1 + X2 + X3) * T",
    family="normal",
    epochs=200,
)
result = model.fit(data)

# Results
print(f"ATE: {result.ate:.3f} (SE: {result.ate_se:.3f})")
print(result.summary())
```

7.3 API Reference

The main estimator class is DeepHTE, which accepts a formula string, family specification, and training hyperparameters. The fit method returns a results object containing the estimated ATE with standard error, individual treatment effects, and methods for prediction on new data.

For simulation studies, the package provides the comparison module with wrappers for EconML's LinearDML and CausalForestDML, as well as QuantileForest for distribution comparison. The simulations module provides organized runners for reproducible Monte Carlo studies across data modalities.

8 Conclusion

DeepHTE provides a principled integration of deep learning with causal inference through the enriched structural models framework. By treating neural network outputs as parameter functions within a well-defined statistical model, the approach achieves the flexibility of deep learning while maintaining the inferential rigor of structural econometrics.

Monte Carlo simulations demonstrate that the method outperforms alternatives with linear CATE assumptions when true heterogeneity is nonlinear. The key advantage is individual treatment effect recovery, which is essential for personalized decision-making. The approach extends naturally to non-tabular covariates through appropriate neural network architectures.

Future work includes extensions to instrumental variables settings, panel data with individual fixed effects, and dynamic treatment regimes. The theoretical framework also admits extensions to more complex structural models beyond treatment effects.

Availability DeepHTE is available at `https://github.com/rawatpranjal/deepstats` under the MIT license. The simulations package provides fully reproducible code for all experiments in this paper.

References

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Farrell, M. H., Liang, T., and Misra, S. (2023). Deep learning for individual heterogeneity: An automatic inference framework. *arXiv preprint arXiv:2010.14694*.