Attention-Based Estimation of Individual Causal Benefit Probabilities under Dose Variation

Lev V. Utkin, Andrei V. Konstantinov, Stanislav K. Kogan, Natalya M. Verbova

Peter the Great St.Petersburg Polytechnic University
Higher School of Artificial Intelligence Technologies

SpacSec'26, October 2026

Authors ...



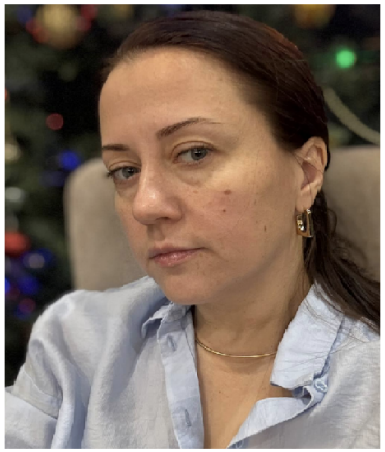
Lev Utkin



Stanislav Kogan



Andrei Konstantinov



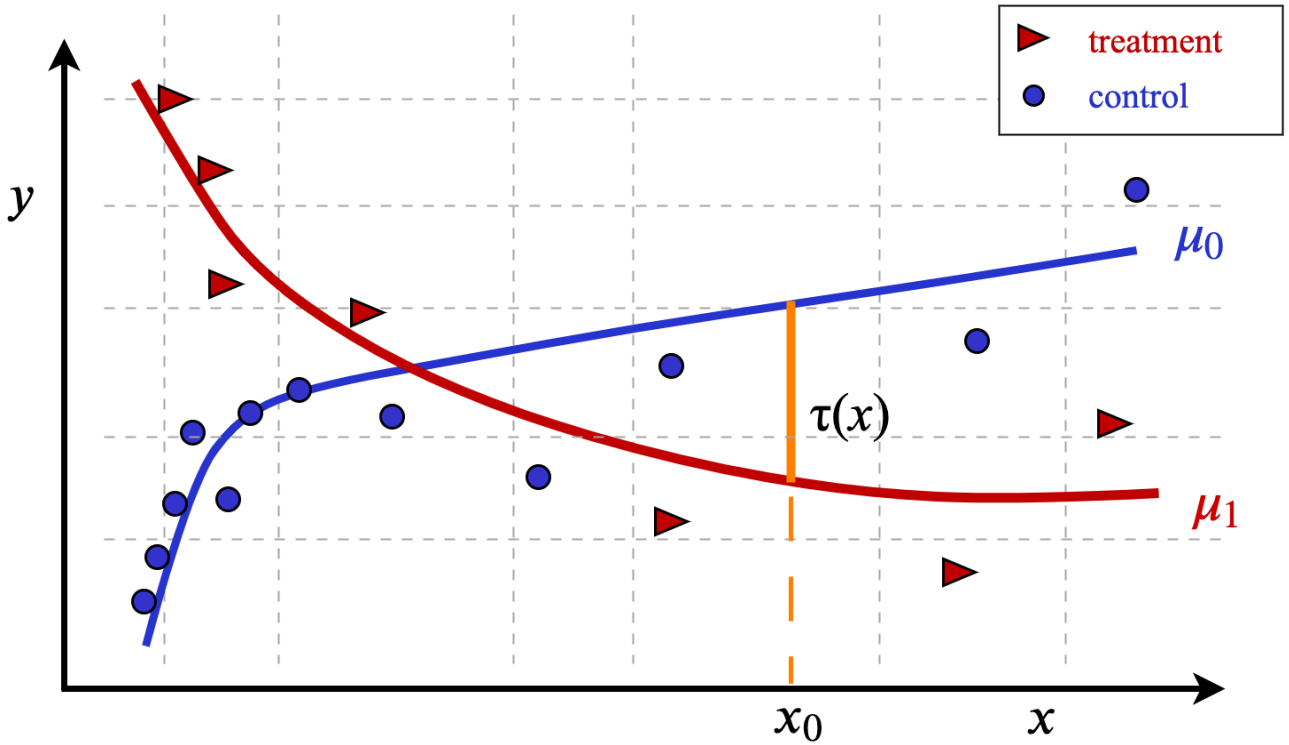
Natalya Verbova

Treatment effect problem statement

- Observations in two groups: control and treatment
- **Control group**: dataset $\mathcal{D}_0 = \{(\mathbf{x}_i, y_i, A_i = 0)\}_{i=1}^{c}$, $y_i \in \mathbb{R}$ is an observation (time until the attack on a server)
- **Treatment group**: dataset $\mathcal{D}_1 = \{(\mathbf{z}_j, h_j, A_j)\}_{j=1}^{t}$, $A_j \in \{1, ..., m\}$ is the treatment dose
- **Goal**: build a model and train it on $\mathcal{D}_0$ and $\mathcal{D}_1$, to estimate the treatment effect for a new unit $\mathbf{x}$
- Individual Probability of Treatment Benefit (**IPTB**):

$$\Psi(\mathbf{x}) = \Pr(H > Y \mid \mathbf{X} = \mathbf{x})$$

Illustration of the Treatment Effect Estimation Problem



Treatment Effect Estimation Problem

- **Main problem**: we never observe an individual treatment effect:
  - For one person, we know either their outcome with treatment or without treatment, but never both simultaneously
  - This is a **fundamental** problem of causal inference (causal effect)
- Consequences of the Fundamental Problem:
  - Treatment is usually not assigned randomly (unprotected units are treated more often).
  - We cannot test the model trained on real data because there are no examples $\mathbf{x}$ with known control y and treatment h outcomes!

Individual Probability of Treatment Benefit (IPTB)

- **IPTB:** $\Psi(\mathbf{x}) = \Pr(H > Y \mid \mathbf{X} = \mathbf{x})$
- We propose a model called the *Dose-AIPTB (Dose Attention-based IPTB):* $A \in \{0, 1, ..., m\}$
- The 1-st idea is the trick:
  - we take $(\mathbf{z}_i, h_i)$ and $(\mathbf{x}_j, y_j)$ from two different groups
  - for each pair, we define the treatment effect (pseudo-effect) $\Delta_{ij} = h_i - y_j$
  - doses are added to the concatenated vectors $\mathbf{z}$ and $\mathbf{x}$ as a feature
- Then our task is to determine the probability

$$p^{+}(\mathbf{z}, \mathbf{x}) = \Pr\{\Delta > 0 \mid \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}\}$$

using all pairs $(\mathbf{z}_i, h_i)$ and $(\mathbf{x}_j, y_j)$.

Treatment doses

- Dose A is added to the concatenated vectors $\mathbf{z}$ and $\mathbf{x}$ as a feature
- It is required weighing features when the dimension of the feature vectors is large
- The weight ω is assigned to the dose feature and weights $(1-\omega)/(2d)$ are assigned to other features
- Distances between feature vectors are calculated in accordance with the weights

The Nadaraya-Watson regression

We can consider our task as a binary classification problem with the class labels $v_{lk} = \mathbf{1}[\Delta_{lk} > 0]$ for every pair:

$$
\begin{aligned}
p^+(\mathbf{z}, \mathbf{x}) &= \sum_{(l,k) \in \mathcal{R}^+ \cup \mathcal{R}^-} w((\mathbf{z}, \mathbf{x}), (\mathbf{z}_l, \mathbf{x}_k)) \cdot v_{lk} \\
&= \sum_{(i,j) \in \mathcal{R}^+} w((\mathbf{z}, \mathbf{x}), (\mathbf{z}_i, \mathbf{x}_j)) \cdot 1 \\
&+ \sum_{(r,s) \in \mathcal{R}^-} w((\mathbf{z}, \mathbf{x}), (\mathbf{z}_r, \mathbf{x}_s)) \cdot 0
\end{aligned}
$$

and the condition

$$
\sum_{(r,s) \in \mathcal{R}^+ \cup \mathcal{R}^-} w((\mathbf{z}, \mathbf{x}), (\mathbf{z}_r, \mathbf{x}_s)) = 1
$$

Attention weights

- Attention weights are expressed through kernels of concatenated pairs of feature vectors:

$$w((\mathbf{z},\mathbf{x}),(\mathbf{z}_i,\mathbf{x}_j)) = \frac{K((\mathbf{z},\mathbf{x}),(\mathbf{z}_i,\mathbf{x}_j))}{\sum_{(r,s)\in\mathcal{R}^+\cup\mathcal{R}^-} K((\mathbf{z},\mathbf{x}),(\mathbf{z}_r,\mathbf{x}_s))}$$

- For the Gaussian kernel:

$$w((\mathbf{z},\mathbf{x}),(\mathbf{z}_i,\mathbf{x}_j)) = \mathrm{softmax}\left(-\frac{\|(\mathbf{z},\mathbf{x})-(\mathbf{z}_i,\mathbf{x}_j)\|^2}{\tau}\right)$$

Model Training

- Loss function is the maximum likelihood

$$\min_{\mathbf{W}_Q, \mathbf{W}_K} \mathcal{L}(\mathbf{p} \mid \mathbf{W}_Q, \mathbf{W}_K) = -\sum_{(i,j)\in\mathcal{R}^+} \log\left(p^+(\mathbf{z}_i, \mathbf{x}_j)\right) - \sum_{(i,j)\in\mathcal{R}^-} \log\left(1 - p^+(\mathbf{z}_i, \mathbf{x}_j)\right) + \eta\left(\|\mathbf{W}_Q\|^2 + \|\mathbf{W}_K\|^2\right)$$

- Codes are publicly available at `https://github.com/NTAILab/AIPTBDose`

Inference

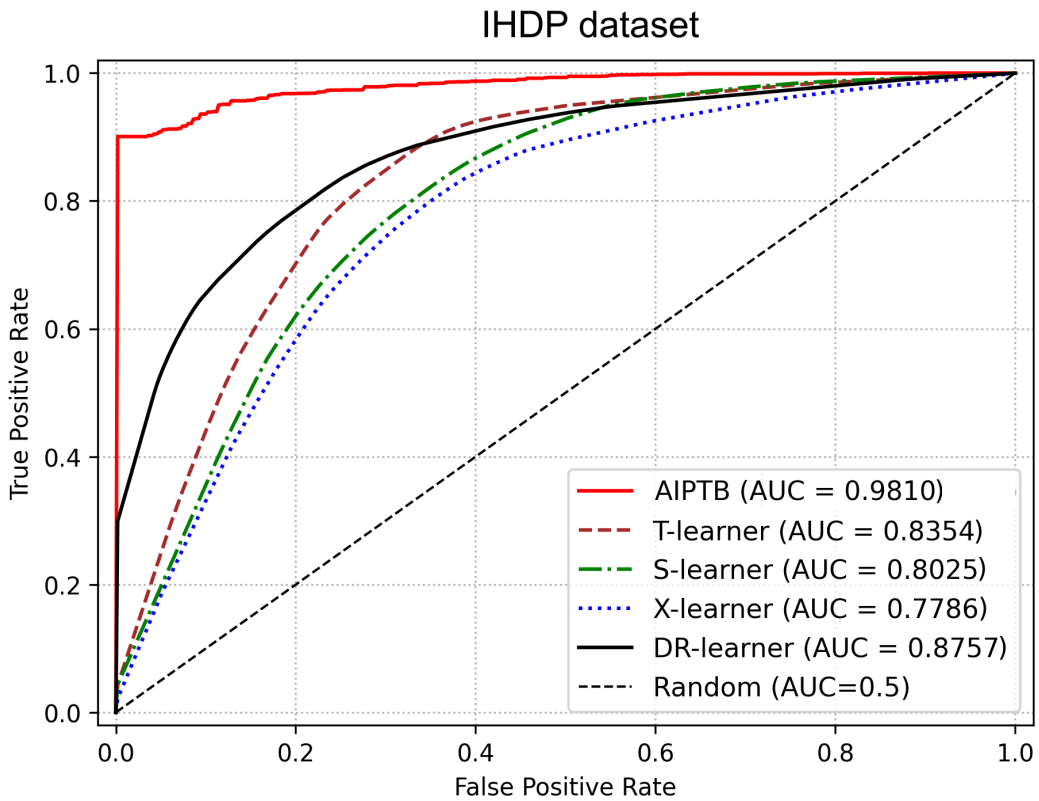
- $p^+(\mathbf{x}, \mathbf{x})$ is computed for a new instance $\mathbf{x}$ under conditions $A = 1, ..., m$
- Suppose that γ is the threshold of the probability $p^+(\mathbf{x}, \mathbf{x})$ for decision making about the positive treatment effect
- Then the number of doses is selected minimizing the difference between $p^+(\mathbf{x}, \mathbf{x})$ and γ by $p^+(\mathbf{x}, \mathbf{x}) \geq \gamma$

Model Features

- The model overcomes the lack of counterfactual data for a unit by forming pairs of observations and estimating the difference between their outcomes Δ_{ij}.
- The approach is effective when the treated group contains few examples and uses all possible pairs of units for training.
- Attention weights allow the model to adaptively assess the contribution of different unit pairs according to their similarity.

Numerical comparison of Dose-AIPTB (ROC-AUC scores)

- Infant Health and Development Program (IHDP) dataset as a benchmark resource frequently employed for CATE estimation



Numerical comparison of Dose-AIPTB trained on synthetic data

Table: AUC values for the models trained on different datasets

Dataset	Dose-AIPTB	T-learner	S-learner	X-learner	DR-learner
Simple	$\mathbf{0.986 \pm 0.011}$	0.767 ± 0.031	0.778 ± 0.036	0.831 ± 0.058	0.847 ± 0.063
Step-wise	$\mathbf{0.844 \pm 0.021}$	0.544 ± 0.031	0.580 ± 0.022	0.811 ± 0.038	0.747 ± 0.052
Linear	$\mathbf{0.954 \pm 0.008}$	0.716 ± 0.021	0.616 ± 0.015	0.802 ± 0.028	0.836 ± 0.049
Spiral	0.843 ± 0.041	0.689 ± 0.019	$\mathbf{0.845 \pm 0.015}$	0.790 ± 0.021	0.723 ± 0.017
Weibull	$\mathbf{0.724 \pm 0.021}$	0.606 ± 0.016	0.575 ± 0.024	0.600 ± 0.021	0.600 ± 0.037
IHDP-100	$\mathbf{0.981 \pm 0.018}$	0.835 ± 0.08	0.802 ± 0.094	0.779 ± 0.104	0.876 ± 0.101

Thank you for your attention

?