PairNet: Training with Observed Pairs to Estimate Individual Treatment Effect

Lokesh Nagalapatti, Pranava Singhal, Avishek Ghosh, Sunita Sarawagi; *Indian Institute of Technology Bombay*

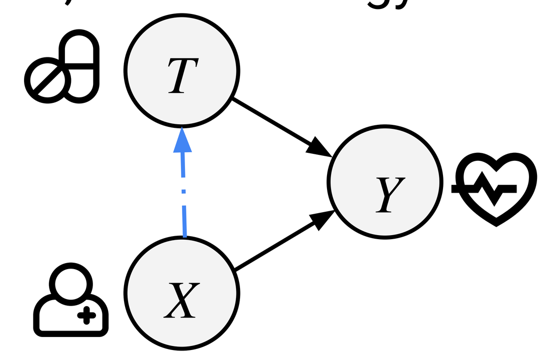
Introduction

Treatment T applied to individual with covariates X leads to outcome $Y(T)$

- **Goal:** estimate Individual Treatment Effect (ITE) $\tau(x, t, t') = \mathbb{E}[Y(t) - Y(t') \mid x]$
- **Challenge:** outcome is observed only under one treatment T. We can not directly regress τ against (X, T, T').

Factual Loss: regress Y against (X, T) to estimate $\mu(x, t) = \mathbb{E}[Y(t) \mid x]$, infer ITE $\hat{\tau}(x, t, t') = \hat{\mu}(x, t) - \hat{\mu}(x, t')$. Only utilises factual outcomes; naive strategy.

- **Confounding**: covariates X are correlated with treatment T in training data. $\hat{\mu}(x, t)$ incurs higher estimation error where $\Pr(t \mid x)$ is low. Pairing each (x, t, y) with (x, t', y') is impossible.



- **Prior works** address this fundamental problem in two broad ways:

Pseudo-outcomes — With
- *Meta-Learners*: two-stage learning, estimate nuisance parameters
- *Matching*: pair (x, t, y) with "nearby" (x', t', y'); assume $\mu(x, t') \approx y'$
- *Generative Models*: model counterfactual distribution

Limitations: faulty pseudo-outcome supervision

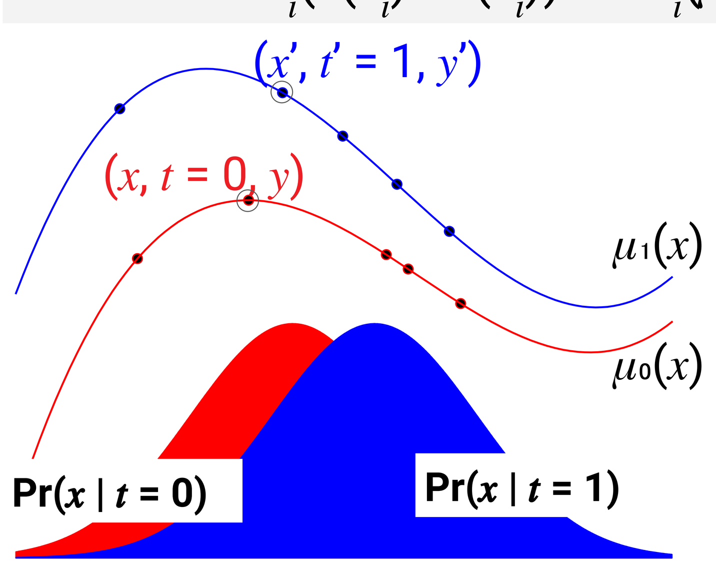
Pseudo-outcomes — Without
- *Regularisation*: balance $\phi(x)$ distributions across treatments
- *Reweighting*: inverse weighting with estimated propensity $\Pr(t \mid x)$

Limitations: lack inductive bias for τ; poor propensity estimates

Motivating Our Approach: PairNet

PairNet avoids pseudo-outcomes by modifying the matching objective.
For binary treatments, $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid x] = \mu_1(x) - \mu_0(x)$

ITE risk = $\Sigma_i(\tau(x_i) - \hat{\tau}(x_i))^2 = \Sigma_i(\mu_1(x_i) - \mu_0(x_i) - \hat{\mu}_1(x_i) + \hat{\mu}_0(x_i))^2$



We can't simultaneously access $\mu_1(x)$ and $\mu_0(x)$ in training data. Pair sample (x, t, y) with nearby sample $(x', 1 - t, y')$

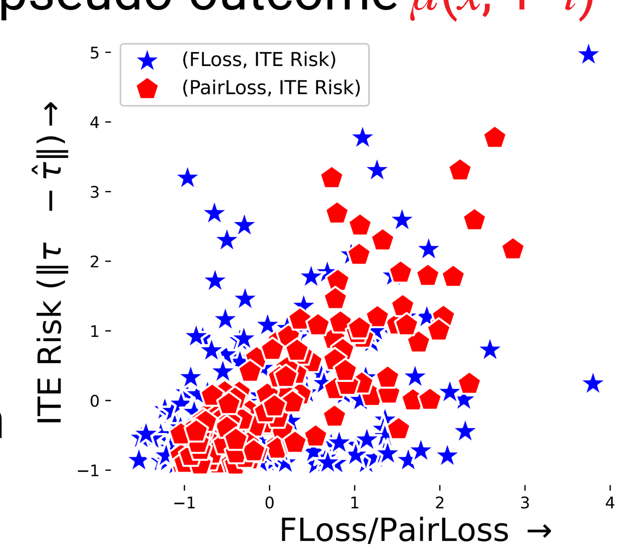
Matching : $\Sigma_i(y - y' - \hat{\mu}(x, t) + \hat{\mu}(x, 1 - t))^2$
PairNet : $\Sigma_i(y - y' - \hat{\mu}(x, t) + \hat{\mu}(x', 1 - t))^2$

PairNet avoids pseudo-outcome $\hat{\mu}(x, 1 - t)$

Pair Loss can be decomposed into factual loss and residual alignment terms:

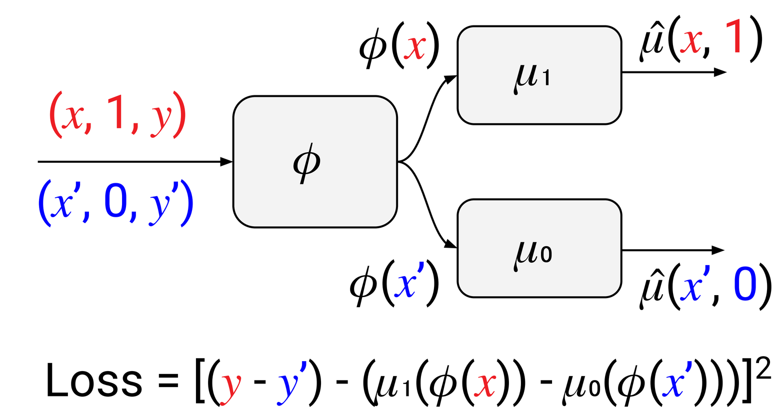
$$\Sigma_i(y - \hat{\mu}(x, t))^2 + (y' - \hat{\mu}(x', 1 - t))^2 - 2\,(y - \hat{\mu}(x, t))(y' - \hat{\mu}(x', 1 - t))$$

The last term promotes a positive correlation among error residuals for near covariates which is a necessary inductive bias for ITE estimation.



PairNet Algorithm

1. Given training data point (x, t, y)
2. Sample alternative treatment t'
3. Sample neighbouring data point (x', t', y') s.t. $d(x, x') = \|\psi(x) - \psi(x')\|$ is small
4. Optimise (ϕ, μ) to minimise Pair Loss



The probability of the j^{th} sample with treatment t', being paired with i^{th} sample with treatment t is proportional to the softmax of the negative distance between them, promoting nearby pairing. This induces a distribution over neighbours:

$$q_t(x' \mid x, t') \propto e^{-d(x, x')} p_{t'}(x') \qquad q_t(x') = \int q_t(x' \mid x, t') p_t(x) dx$$

Theoretical Analysis: Bounds on ITE Risk for Binary Treatment

Define the error residue $r_t(x) = \hat{\mu}(x, t) - \mu(x, t)$; $u_t = \Pr(t)$; $p_t(x) = \Pr(x \mid t)$

ITE risk $\varepsilon_{ITE} = \int_x (r_1(x) - r_0(x))^2 p(x) dx$

$= \Sigma_t u_t [\int_x r_t(x)^2 p_t(x) dx + \int_x r_{1-t}(x)^2 p_t(x) dx - 2 \int_x r_t(x) r_{1-t}(x) p_t(x) dx]$

Pair Loss $\varepsilon_{pair} = \Sigma_t u_t \int_x \int_{x'} (r_t(x) - r_{1-t}(x'))^2 p_t(x)\, q_t(x' \mid x) dx' dx$

$= \Sigma_t u_t [\int_x r_t(x)^2 p_t(x) dx + \int_{x'} r_{1-t}(x')^2 q_t(x') dx' - 2 \int_x \int_{x'} r_t(x) r_{1-t}(x') p_t(x)\, q_t(x' \mid x) dx' dx]$

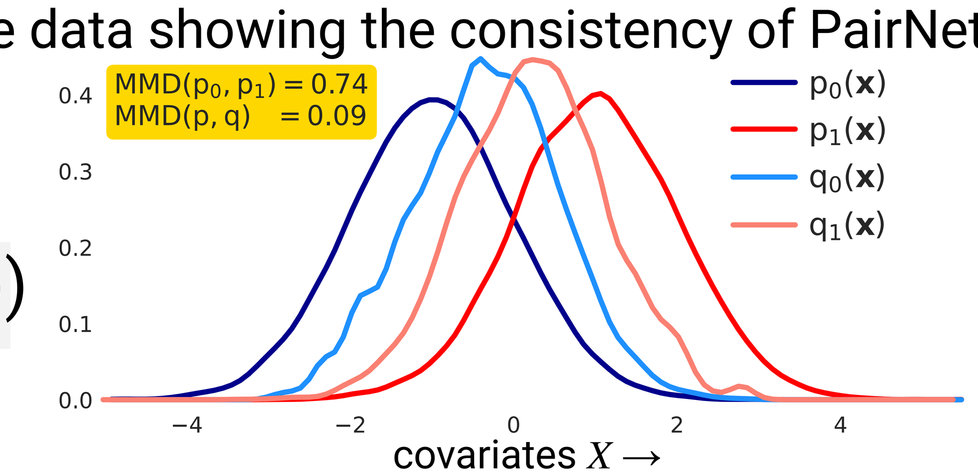
Integral Probability Metric $IPM_G(p, q) = \sup_{g \in G} |\int g(x)(p(x) - q(x)) dx|$

We show that $\varepsilon_{ITE} \leq \varepsilon_{pair} + \Sigma_t u_t \,[\, B \cdot IPM_G(p_t, q_t) + 2\, K_{1-t}\, \delta \sqrt{\varepsilon^t_F}\,]$

assuming expected neighbour distance $\leq \delta$, r_t is K_t-Lipschitz and $r_t^2/B \in G$.
The bound converges to zero for large data showing the consistency of PairNet.

This bound is tighter than the bound of Shalit et al. for Factual loss:

$$\varepsilon_{ITE} \leq 2(\varepsilon^0_F + \varepsilon^1_F + B \cdot IPM_G(p_0, p_1))$$

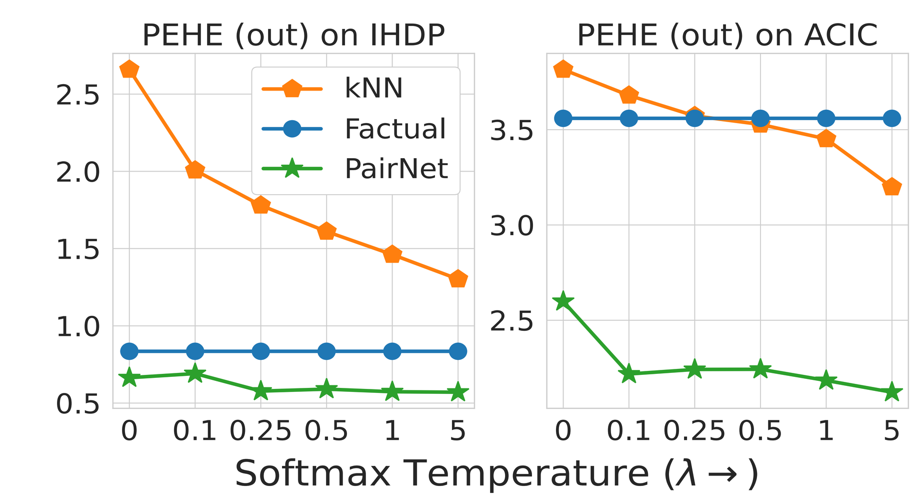


Experiments

- **Performance Metric:** PEHE error (square root of empirical ITE risk); we also report p-values for a one-sided **paired t-test** comparing PairNet to baselines
- **Datasets:** IHDP, ACIC, and Twins (Binary); TCGA[0-2], IHDP, News (Continuous)
- **PairNet is model agnostic**; can use any T-Learner architecture. We use **TARNet**
- For continuous treatments we consider both **DRNet** and **VCNet** architectures
- Implemented in **JAX** within the **CATENets library** with default hyperparameters
- PairNet constructs pairs using ψ, the representation ϕ trained on factual loss.
- Hyperparameters: δ_{pair} (fraction far pairs dropped) and $num_{z'}$ (# pairs/sample)

- PairNet outperforms state-of-the-art ITE estimators across binary and continuous treatment benchmarks with high statistical significance

	Estimator	IHDP	ACIC	Twins
Meta Learners	TLearner	1.34 (0.00)	4.29 (0.03)	0.32 (0.01)
	RLearner	3.24 (0.00)	3.94 (0.00)	0.32 (0.15)
	DRLearner	1.35 (0.00)	3.33 (0.08)	0.32 (0.14)
	XLearner	1.91 (0.00)	3.31 (0.10)	0.32 (0.01)
Representation Learners	TARNet	0.83 (0.11)	2.71 (0.29)	0.32 (0.00)
	CFRNet	1.11 (0.00)	3.45 (0.06)	0.33 (0.00)
	FlexTENet	1.26 (0.00)	5.37 (0.00)	0.36 (0.00)
Weighting	IPW	0.93 (0.04)	2.57 (0.41)	0.33 (0.00)
	DragonNet	0.83 (0.11)	2.72 (0.28)	0.33 (0.00)
	PairNet	0.69 (0.00)	2.46 (0.00)	0.32 (0.00)



- PairNet outperforms matching (kNN) and Factual across different levels of proximity between covariates in a pair
- PairNet is less sensitive to variation in proximity, outperforming factual loss even for random pairs

- To create pairs for continuous treatments first sample treatment $t^0 \sim U(0, 1)$
- Then we sample (x', t', y') such that $|t^0 - t'| < 0.05$
- PairNet outperforms VCNet significantly when dataset size is reduced

	IHDP	News	TCGA-0 drop 90% data
DRNet	2.45 (0.00)	1.42 (0.00)	0.52 (0.00)
PairNet	2.27(0.00)	1.32(0.00)	0.44 (0.00)
VCNet	1.73 (0.02)	1.24(1.00)	0.43 (0.02)
PairNet	1.57(0.00)	1.26 (0.00)	0.27 (0.00)

- PairNet is not very sensitive to hyperparameters δ_{pair} and $num_{z'}$
- When applying Pair Loss to other representation learning T-Learners (CFRNet, DragonNet, FlexTENet) we observe similar performance gains
- We do not observe any statistically significant variation in performance on changing the weight of the residue alignment term $(y - \hat{\mu}(x, t))(y' - \hat{\mu}(x', 1 - t))$

Scan QR code to access full paper, code and author homepages