LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

Normalizing Flows for Interventional Density Estimation

Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel

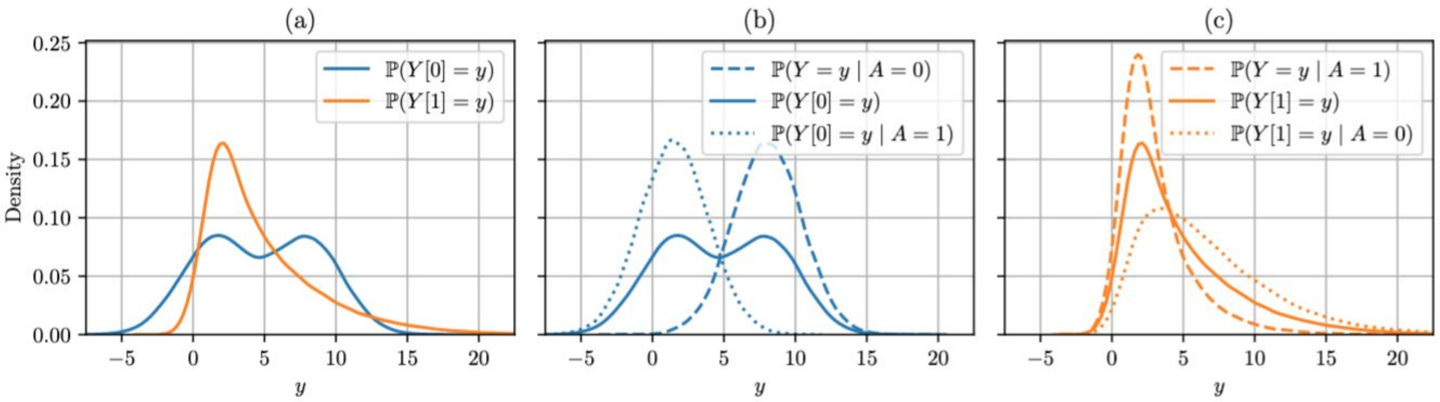
LMU Munich & Munich Center for Machine Learning (MCML), Munich, Germany

ICML 2023, Short Presentation

Introduction: Efficient interventional density estimation

Why this is important?

- Making decisions based on averaged causal quantities can be misleading and, in some applications, even dangerous



$$\mathbb{E}(Y[0]) = \mathbb{E}(Y[1]) \approx 4.77$$
$$\mathrm{var}(Y[0]) = \mathrm{var}(Y[1]) \approx 4.06$$

$$\mathbb{P}\{Y[1] < 5.0\} \approx 0.63 \quad \mathbb{P}\{Y[0] < 5.0\} \approx 0.51$$

Problem formulation

Given observational dataset of:

X covariates
A binary treatments
Y continuous (factual) outcomes

we want to flexibly and efficiently estimate **interventional density** (density of the potential outcomes)

$$\mathbb{P}(Y[a] = y) = \mathop{\mathbb{E}}_{X \sim \mathbb{P}(X)} \left(\mathbb{P}(Y = y \mid X, A = a)\right)$$

Introduction: Task complexity – Assumptions

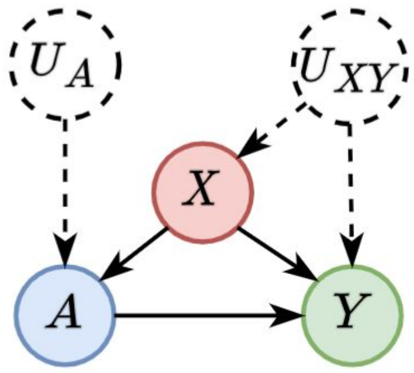
Why estimation is hard?

- Traditional density estimation is non-applicable for **Interventional Density Estimation** (IDE), as we do not have samples from interventional distributions (i.e., the fundamental problem of causal inference)
- Density is a **functional, infinitely-dimensional target estimand**, and, hence, standard semi-parametric efficiency theory (with influence functions) is not applicable.
- Choice of the nuisance parameters on practice: conditional expectations vs. conditional densities?

Identifiability assumptions

Potential outcomes framework (Neuman-Rubin)[1]

- **Consistency**. If $A = a$ is a treatment for some patient, then $Y = Y[a]$
- **Positivity (Overlap).** There is always a non-zero probability of receiving/not receiving any treatment, conditioning on the covariates: $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$
- **Exchangeability (Ignorability).** Current treatment is independent of the potential outcome, conditioning on the covariates $A \perp\!\!\!\perp Y[a] \mid X$ for all a.



[1] Neyman, J. S. On the application of probability theory to agricultural experiments. Annals of Agricultural Sciences, 10:1–51, 1923.
Rubin, D. B. Bayesian inference for causal effects: The role of randomization. The Annals of Statistics, pp. 34–58, 1978.

Introduction: Related work - Research gap – Our contributions

Related methods

Method	Parametric	Estimator type	Efficiency wrt.	Base density model	Proper density	Universal
Kim et al. (2018)	semi-parametric	A-IPTW	L_1 distance	kernel density estimation (KDE)	✗	✓
Muandet et al. (2021)	non-parametric	plug-in	—	distributional kernel mean embeddings (DKME)	✗	✓
Kennedy et al. (2023)	semi- / fully-parametric	A-IPTW	moment condition	exponential family	✓	✗
				truncated series (TS)	✗	✓
INFs (this paper)	fully-parametric	A-IPTW	moment condition	normalizing flows (NFs)	✓	✓

A-IPTW: augmented inverse propensity of treatment weighted

Research gap

- Existing methods for IDE are either non- or semi-parametric. Our work is the first to propose a **universal fully-parametric**, deep learning method for IDE, with proper density.

Our contributions

Interventional Normalizing Flows (INFs) are first proper fully-parametric, deep learning method for interventional density estimation:

- We extend the results of (Kennedy et al., 2023)[1] and derive a tractable optimization problem with a one-step bias correction for efficient and doubly robust estimation. This allows for an effective two-step training procedure.
- We demonstrate in various experiments that INFs are highly expressive and effective. A major advantage owed to the parametric form is that our INFs scale well to both large and high-dimensional datasets.

[1] Kennedy, E. H., Balakrishnan, S., and Wasserman, L. Semi-parametric counterfactual density estimation. Biometrika, 2023.

INFs: Semi-parametric interventional density estimation (IDE) (Kennedy et al., 2023)[1]

Target: interventional density $\mathbb{P}(Y[a] = y) = \mathop{\mathbb{E}}_{X \sim \mathbb{P}(X)} \big(\mathbb{P}(Y = y \mid X, A = a)\big)$

One-step IDE

Two-step IDE

[1] Kennedy, E. H., Balakrishnan, S., and Wasserman, L. Semi-parametric counterfactual density estimation. Biometrika, 2023.

INFs: Semi-parametric interventional density estimation (IDE) (Kennedy et al., 2023)[1]

One-step IDE

Target: interventional density $\mathbb{P}(Y[a] = y) = \underset{X \sim \mathbb{P}(X)}{\mathbb{E}} \big(\mathbb{P}(Y = y \mid X, A = a)\big)$

- **Plug-in estimator**:

  $$\hat{\mathbb{P}}^{\mathrm{PI}}(Y[a] = y) = \mathbb{P}_n\{\hat{\mathbb{P}}(Y = y \mid X, A = a)\}.$$

Two-step IDE

[1] Kennedy, E. H., Balakrishnan, S., and Wasserman, L. Semi-parametric counterfactual density estimation. Biometrika, 2023.

INFs: Semi-parametric interventional density estimation (IDE) (Kennedy et al., 2023)[1]

One-step IDE

Target: interventional density $\mathbb{P}(Y[a] = y) = \underset{X \sim \mathbb{P}(X)}{\mathbb{E}} \big(\mathbb{P}(Y = y \mid X, A = a)\big)$

- **Plug-in estimator**:

$$\hat{\mathbb{P}}^{\mathrm{PI}}(Y[a] = y) = \mathbb{P}_n\{\hat{\mathbb{P}}(Y = y \mid X, A = a)\}.$$

Target: projection parameters $\hat{\beta}_a = \arg\min_{\beta_a} \mathrm{KL}\left(\mathbb{P}(Y[a]) \,\|\, g(\cdot; \beta_a)\right) = \arg\min_{\beta_a} \underset{Y^a \sim \mathbb{P}(Y[a])}{\mathbb{E}} \left(-\log g(Y^a; \beta_a)\right).$

$\iff$ solving a moment equation $m(\beta_a) = \underset{Y^a \sim \mathbb{P}(Y[a])}{\mathbb{E}} T(Y^a; \beta_a) \overset{!}{=} 0$

Score function:
$T(Y; \beta_a) = -\nabla_{\beta_a} \log g(Y; \beta_a)$

Two-step IDE

[1] Kennedy, E. H., Balakrishnan, S., and Wasserman, L. Semi-parametric counterfactual density estimation. Biometrika, 2023.

INFs: Semi-parametric interventional density estimation (IDE) (Kennedy et al., 2023)[1]

One-step IDE

Target: interventional density $\mathbb{P}(Y[a] = y) = \mathop{\mathbb{E}}_{X \sim \mathbb{P}(X)} \big(\mathbb{P}(Y = y \mid X, A = a)\big)$

- **Plug-in estimator**:

$$\hat{\mathbb{P}}^{\mathrm{PI}}(Y[a] = y) = \mathbb{P}_n\{\hat{\mathbb{P}}(Y = y \mid X, A = a)\}.$$

Two-step IDE

Target: projection parameters $\hat{\beta}_a = \arg\min_{\beta_a} \mathrm{KL}\,\big(\mathbb{P}(Y[a]) \,\|\, g(\cdot; \beta_a)\big) = \arg\min_{\beta_a} \mathop{\mathbb{E}}_{Y^a \sim \mathbb{P}(Y[a])} \big(-\log g(Y^a; \beta_a)\big).$

$\iff$ solving a moment equation $m(\beta_a) = \mathop{\mathbb{E}}_{Y^a \sim \mathbb{P}(Y[a])} T(Y^a; \beta_a) \stackrel{!}{=} 0$

Score function: $T(Y; \beta_a) = -\nabla_{\beta_a} \log g(Y; \beta_a)$

- **Covariate-adjusted estimator:**

$$\hat{\mathbb{P}}^{\mathrm{CA}}(Y[a] = y) = g(y; \hat{\beta}_a^{\mathrm{PI}}) \qquad \hat{m}^{\mathrm{PI}}(\beta_a) = \mathop{\mathbb{E}}_{Y^a \sim \mathbb{P}_n\{\hat{\mathbb{P}}(Y \mid X, A=a)\}} T(Y^a; \beta_a) \stackrel{!}{=} 0$$

[1] Kennedy, E. H., Balakrishnan, S., and Wasserman, L. Semi-parametric counterfactual density estimation. Biometrika, 2023.

INFs: Semi-parametric interventional density estimation (IDE) (Kennedy et al., 2023)[1]

One-step IDE

Target: interventional density $\mathbb{P}(Y[a]=y) = \mathop{\mathbb{E}}_{X\sim\mathbb{P}(X)}\big(\mathbb{P}(Y=y \mid X, A=a)\big)$

- **Plug-in estimator**:

$$\hat{\mathbb{P}}^{\mathrm{PI}}(Y[a]=y) = \mathbb{P}_n\{\hat{\mathbb{P}}(Y=y \mid X, A=a)\}.$$

Two-step IDE

Target: projection parameters $\hat{\beta}_a = \arg\min_{\beta_a} \mathrm{KL}\left(\mathbb{P}(Y[a]) \,\|\, g(\cdot;\beta_a)\right) = \arg\min_{\beta_a} \mathop{\mathbb{E}}_{Y^a\sim\mathbb{P}(Y[a])}\left(-\log g(Y^a;\beta_a)\right).$

$\iff$ solving a moment equation $m(\beta_a) = \mathop{\mathbb{E}}_{Y^a\sim\mathbb{P}(Y[a])} T(Y^a;\beta_a) \overset{!}{=} 0$

Score function: $T(Y;\beta_a) = -\nabla_{\beta_a}\log g(Y;\beta_a)$

- **Covariate-adjusted estimator:**

$$\hat{\mathbb{P}}^{\mathrm{CA}}(Y[a]=y) = g(y;\hat{\beta}_a^{\mathrm{PI}}) \qquad \hat{m}^{\mathrm{PI}}(\beta_a) = \mathop{\mathbb{E}}_{Y^a\sim\mathbb{P}_n\{\hat{\mathbb{P}}(Y\mid X, A=a)\}} T(Y^a;\beta_a) \overset{!}{=} 0$$

- **Augmented inverse propensity of treatment weighted (A-IPTW) estimator:**

$$\hat{\mathbb{P}}^{\text{A-IPTW}}(Y[a]=y) = g(y;\hat{\beta}_a^{\text{A-IPTW}}) \qquad \hat{m}^{\text{A-IPTW}}(\beta_a) = \hat{m}^{\mathrm{PI}}(\beta_a) + \mathbb{P}_n\{\phi_a(T(Y;\beta_a);\hat{\mathbb{P}})\} \overset{!}{=} 0.$$

efficient influence function: $\phi_a(T;\mathbb{P}) = \dfrac{\mathbb{1}(A=a)}{\pi_a(X)}\big(T - \mathbb{E}(T \mid X, A=a)\big) + \mathbb{E}(T \mid X, A=a) - \mathop{\mathbb{E}}_{X\sim\mathbb{P}(X)}\big(\mathbb{E}(T \mid X, A=a)\big)$

[1] Kennedy, E. H., Balakrishnan, S., and Wasserman, L. Semi-parametric counterfactual density estimation. Biometrika, 2023.

INFs: Semi-parametric interventional density estimation (IDE) (Kennedy et al., 2023)[1]

One-step IDE

Target: interventional density $\mathbb{P}(Y[a]=y) = \underset{X\sim\mathbb{P}(X)}{\mathbb{E}}\big(\mathbb{P}(Y=y \mid X, A=a)\big)$

- **Plug-in estimator**:

$$\hat{\mathbb{P}}^{\mathrm{PI}}(Y[a]=y) = \mathbb{P}_n\{\hat{\mathbb{P}}(Y=y \mid X, A=a)\}.$$

Two-step IDE

Target: projection parameters $\hat{\beta}_a = \arg\min_{\beta_a} \mathrm{KL}\big(\mathbb{P}(Y[a]) \,\|\, g(\cdot;\beta_a)\big) = \arg\min_{\beta_a} \underset{Y^a\sim\mathbb{P}(Y[a])}{\mathbb{E}}\big(-\log g(Y^a;\beta_a)\big)$

$\iff$ solving a moment equation $m(\beta_a) = \underset{Y^a\sim\mathbb{P}(Y[a])}{\mathbb{E}} T(Y^a;\beta_a) \overset{!}{=} 0$

Score function: $T(Y;\beta_a) = -\nabla_{\beta_a}\log g(Y;\beta_a)$

- **Covariate-adjusted estimator:**

$$\hat{\mathbb{P}}^{\mathrm{CA}}(Y[a]=y) = g(y;\hat{\beta}_a^{\mathrm{PI}}) \qquad \hat{m}^{\mathrm{PI}}(\beta_a) = \underset{Y^a\sim\mathbb{P}_n\{\hat{\mathbb{P}}(Y\mid X, A=a)\}}{\mathbb{E}} T(Y^a;\beta_a) \overset{!}{=} 0$$

- **Augmented inverse propensity of treatment weighted (A-IPTW) estimator:**

$$\hat{\mathbb{P}}^{\text{A-IPTW}}(Y[a]=y) = g(y;\hat{\beta}_a^{\text{A-IPTW}}) \qquad \hat{m}^{\text{A-IPTW}}(\beta_a) = \hat{m}^{\mathrm{PI}}(\beta_a) + \mathbb{P}_n\{\phi_a(T(Y;\beta_a);\hat{\mathbb{P}})\} \overset{!}{=} 0.$$

efficient influence function: $\phi_a(T;\mathbb{P}) = \dfrac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(T - \mathbb{E}(T\mid X, A=a)\Big) + \mathbb{E}(T\mid X, A=a) - \underset{X\sim\mathbb{P}(X)}{\mathbb{E}}\big(\mathbb{E}(T\mid X, A=a)\big)$

[1] Kennedy, E. H., Balakrishnan, S., and Wasserman, L. Semi-parametric counterfactual density estimation. Biometrika, 2023.

INFs: Novel efficient optimization objective

Proposed by (Kennedy et al., 2023)[1]

A-IPTW estimator = solution of the multivariate system of equations:

$$\hat{m}^{\text{A-IPTW}}(\beta_a) = \hat{m}^{\text{PI}}(\beta_a) + \mathbb{P}_n\{\phi_a(T(Y;\beta_a);\hat{\mathbb{P}})\} \stackrel{!}{=} 0.$$

$$\phi_a(T;\mathbb{P}) = \frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(T - \mathbb{E}(T \mid X, A=a)\Big) + \mathbb{E}(T \mid X, A=a) - \mathop{\mathbb{E}}_{X\sim\mathbb{P}(X)}(\mathbb{E}(T \mid X, A=a)).$$

Our idea

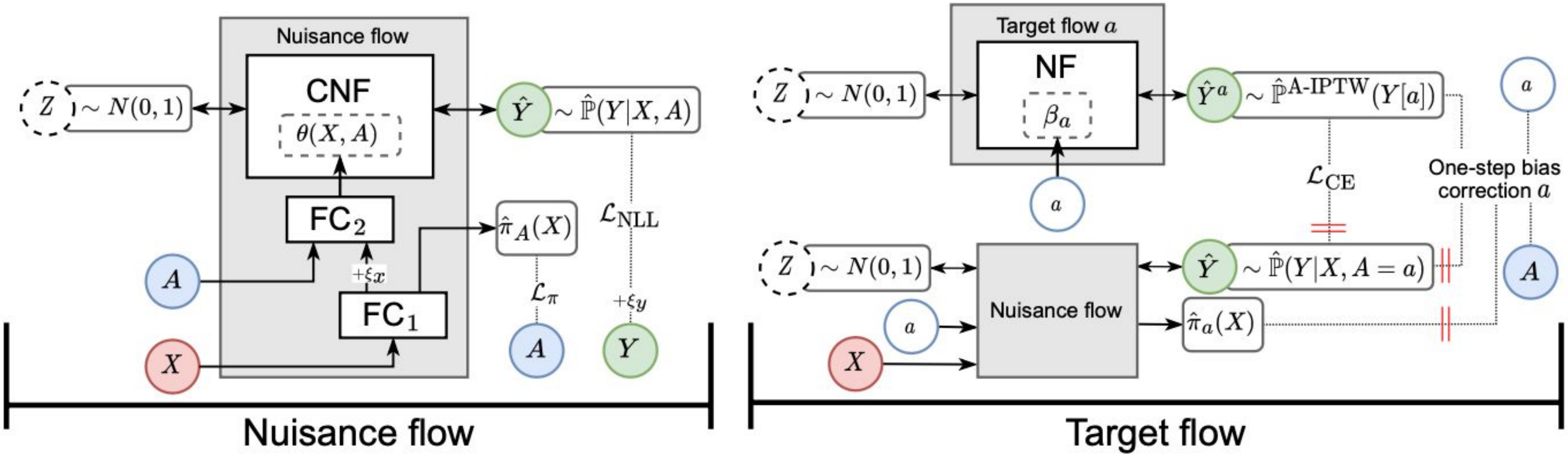
A-IPTW estimator = solution of the optimization task:

$$\hat{\beta}_a^{\text{A-IPTW}} = \arg\min_{\beta_a}\Bigg[\underbrace{\mathop{\mathbb{E}}_{Y^a\sim\mathbb{P}_n\{\hat{\mathbb{P}}(Y\mid X, A=a)\}}\Big(-\log g(Y^a;\beta_a)\Big)}_{\text{cross-entropy loss}}$$

$$\underbrace{-\,\mathbb{P}_n\left\{\frac{\mathbb{1}(A=a)}{\hat{\pi}_a(X)}\Big(\log g(Y;\beta_a) - \mathop{\mathbb{E}}_{Y\sim\hat{\mathbb{P}}(Y\mid X, A=a)}\big(\log g(Y;\beta_a)\big)\Big)\right\}}_{\text{one-step bias correction}}\Bigg]$$

[1] Kennedy, E. H., Balakrishnan, S., and Wasserman, L. Semi-parametric counterfactual density estimation. Biometrika, 2023.

INFs: Novel architecture – Losses



$$\mathcal{L}_{\text{NLL}} \leftarrow -\log \hat{\mathbb{P}}(Y = \tilde{Y} \mid X, A)$$
$$\mathcal{L}_{\pi} \leftarrow \text{BCE}(\hat{\pi}_A(X), A)$$
$$\mathcal{L}_{\text{N}}(\hat{\mathbb{P}}, \hat{\pi}_a) \leftarrow \mathbb{P}^{\mathcal{B}}_{b_{\text{N}}}\{\mathcal{L}_{\text{NLL}} + \alpha \mathcal{L}_\pi\}$$

$$\mathcal{L}_{\text{CE}}(\beta_a^{(i)}) \leftarrow -h \sum_{j=1}^{K} \log g(y_j; \beta_a^{(i)}) \mathbb{P}^{\mathcal{B}}_{b_{\text{T}}}\{\hat{\mathbb{P}}(Y = y_j \mid X, A = a)\}$$
$$\mathcal{L}_{\text{CCE}}(X; \beta_a^{(i)}) \leftarrow -h \sum_{j=1}^{K} \log g(y_j; \beta_a^{(i)}) \hat{\mathbb{P}}(Y = y_j \mid X, A = a)$$
$$\text{bias correction}(\beta_a^{(i)}) \leftarrow \mathbb{P}^{\mathcal{B}}_{b_{\text{T}}}\left\{ \frac{\mathbb{1}(A = a \,\&\, \hat{\pi}_a(X) \geq 0.05)}{\hat{\pi}_a(X)} \Big(-\log g(Y; \beta_a^{(i)}) - \mathcal{L}_{\text{CCE}}(X; \beta_a^{(i)}) \Big) \right\}$$
$$\mathcal{L}_{\text{T}}(\beta_a^{(i)}) \leftarrow \mathcal{L}_{\text{CE}}(\beta_a^{(i)}) + \text{bias correction}(\beta_a^{(i)})$$

Experiments: Datasets – Results

Datasets

- We evaluate INFs based on 1 synthetic, 77 + 24 + 2 semi-synthetic and 1 real-world datasets
- Only synthetic and semi-synthetic data have ground-truth potential outcomes; real-world evaluation is a proof of concept
- We compared test log-probability for each potential outcome (higher is better)

Results

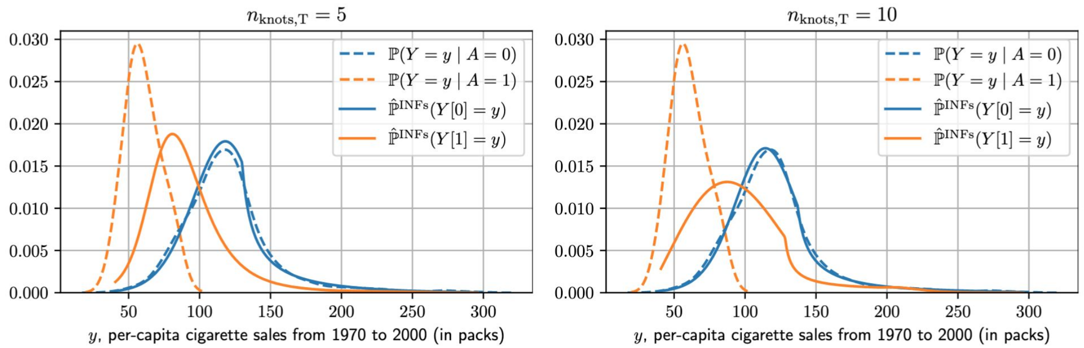
INFs achieve **superior performance** and scales well:

ACIC 2016 & 2018 datasets

	ACIC 2016 (77 datasets)		ACIC 2018 (24 datasets)	
	% best_{in}	% best_{out}	% best_{in}	% best_{out}
TARNet*	3.90%	6.23%	7.08%	7.50%
MDNs	28.96%	29.35%	21.25%	18.75%
CNF [$\hat{=}$ INFs w/o target flow]	14.42%	15.97%	14.17%	14.58%
KDE (Kim et al., 2018)	1.04%	1.04%	10.42%	9.58%
DKME (Muandet et al., 2021)	0.39%	0.78%	8.75%	10.83%
CNF+TS (Kennedy et al., 2023)	8.18%	8.96%	5.83%	5.42%
INFs w/o bias corr	5.45%	7.27%	4.58%	5.42%
INFs (main)	**37.66%**	**30.39%**	**27.92%**	**27.92%**

Higher = better (best in bold)

California's Tobacco Control Program



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
LMU MUNICH SCHOOL OF MANAGEMENT
INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Conclusion

For decision-making in personalized medicine, it is not only important to know **how likely it is that treatments achieve the desired outcome**.

To address this, we propose a novel method for **estimating the density of potential outcomes**. Specifically, we present our **Interventional Normalizing Flows**, which is the **first, fully-parametric, deep learning method** for this purpose.

Source Code:
github.com/Valentyn1997/INFs

ArXiv Paper:
arxiv.org/abs/2209.06203