Reliable Off-Policy Learning for Dosage Combinations

Jonas Schweisthal, Dennis Frauen, Valentyn Melnychuk, Stefan Feuerriegel

Institute of AI in Management
LMU Munich
https://www.som.lmu.de/ai/

Motivation: Real-world interest in finding optimal dosage combinations

Dosage combinations in medicine

- Multiple simultaneously assigned treatments with dosages
- Joint effects (drug-drug interactions)
- Applications: combination therapies for cancer, mechanical ventilation at ICU

Goal

- Finding optimal individualized dosage combinations

Constraints

- Accurate dose-response estimation taking joint effects into account
- Scalability
- Reliable dosing recommendations

➡ **Tackle the challenge with causal inference**

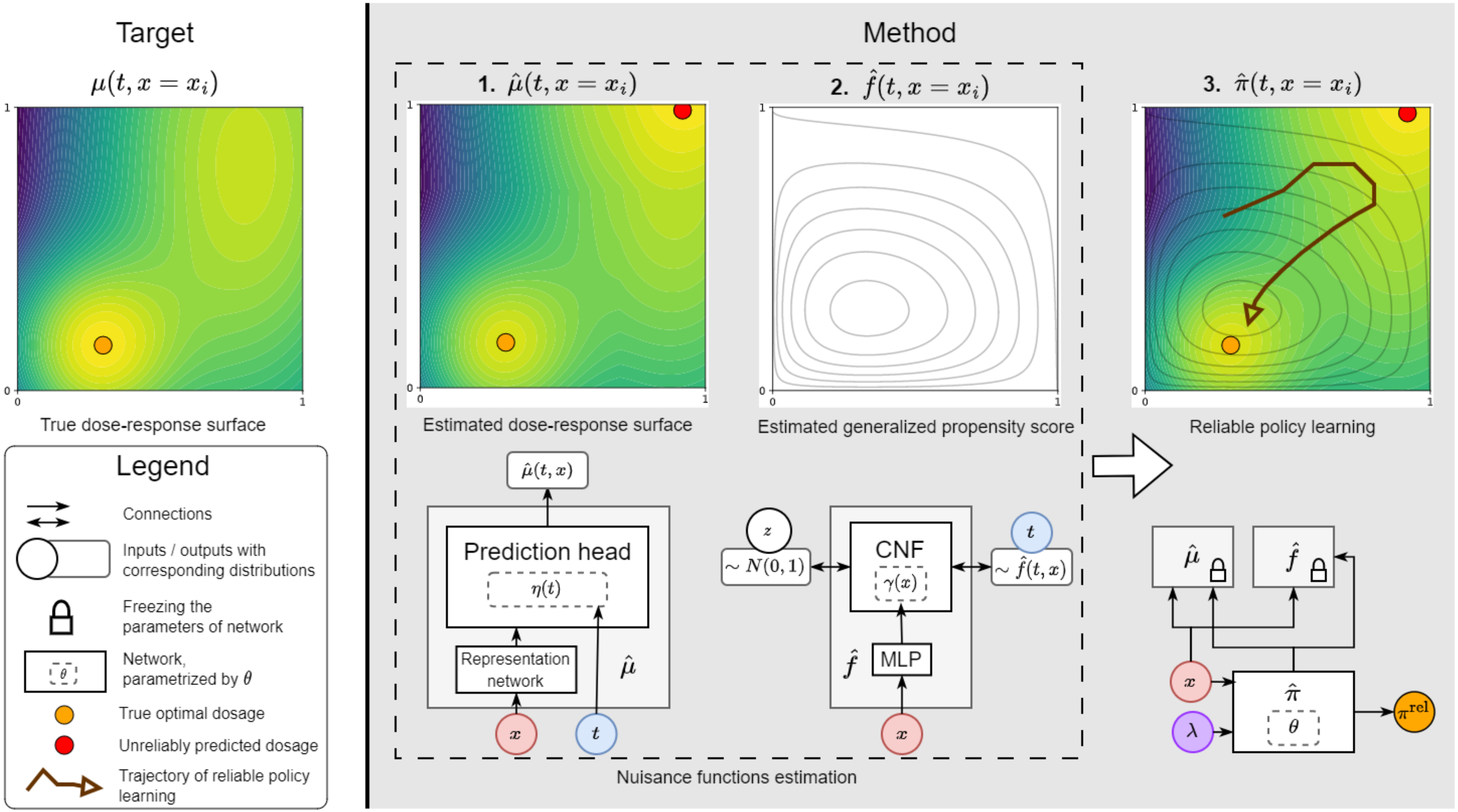
Inherent challenges

1. Observational data

2. Modeling the joint effect of multiple continuous treatments

3. Areas with limited overlap (low data support) in the covariate-treatment space

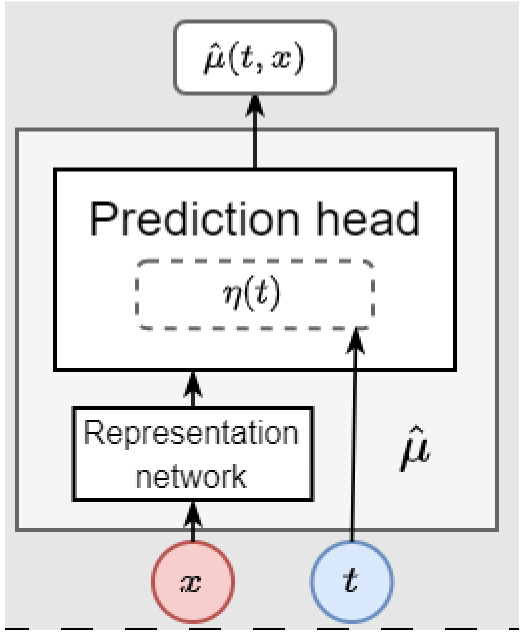
Method: Reliable Off-Policy Learning for Dosage Combinations



1. DCNet for estimating the individualized dose-response function

- Representation network for learning representations of high-dimensional patient features
- Tailored prediction head
  - Expressiveness: Influence of dosages not lost in high-dimensional feature space
  - Smoothness: Model non-linear dose-response surface across treatments smoothly

$$\eta_j(t) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \cdots \sum_{k_p=1}^{K_p} \beta_{j,k_1 k_2 \dots k_p} \cdot \psi_{k_1}^{(1)}(t_1) \cdot \psi_{k_2}^{(2)}(t_2) \cdot \dots \cdot \psi_{k_p}^{(p)}(t_p),$$

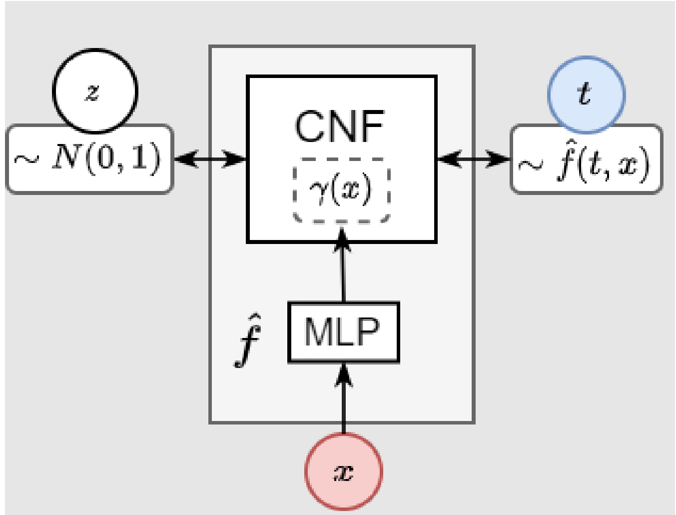


2. Conditional normalizing flows for estimating the GPS

- Neural spline flows for flexible estimation of multimodal densities
- Masked autoregressive networks for multidimensional conditional density estimation

Advantages:

- Universal density approximators
- Properly normalized
- After training constant inference time



3. Reliable Policy Learning

- Objective

$$\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}\left(\pi_\theta(x_i), x_i\right) \quad \text{s.t. } \hat{f}\left(\pi_\theta(x_i), x_i\right) \geq \bar{\varepsilon}, \forall i$$

$$\Longleftrightarrow \min_{\theta}\max_{\lambda_i \geq 0} -\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}\left(\pi_\theta(x_i), x_i\right) - \lambda_i\left[\hat{f}\left(\pi_\theta(x_i), x_i\right) - \bar{\varepsilon}\right]\right\}$$

- Can be solved by gradient descent-ascent wrt. to train loss

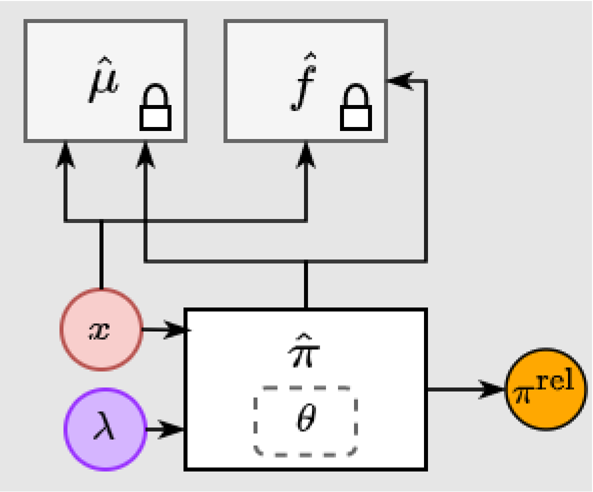
$$\mathcal{L}_\pi(\theta, \lambda) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}\left(\hat{\pi}_\theta(x_i), x_i\right) - \lambda_i\left[\hat{f}\left(\hat{\pi}_\theta(x_i), x_i\right) - \bar{\varepsilon}\right]\right\}$$

- No guarantee for global convergence in the non-convex optimization problem → select best out of k runs wrt. validation loss

$$\pi_\theta^{\text{rel}} = \pi_\theta^{(j)}, \quad \text{with} \quad j = \arg\max_{k} \sum_{i=1}^{n}\hat{\mu}\left(\pi_\theta^{(k)}(x_i), x_i\right) \cdot \mathbb{1}\left\{\hat{f}\left(\pi_\theta^{(k)}(x_i), x_i\right) \geq \bar{\varepsilon}\right\}$$

Algorithm 1: Reliable off-policy learning for dosage combinations

Input : data (X, T, Y), reliability threshold $\bar{\varepsilon}$
Output : optimal reliable policy $\hat{\pi}_\theta^{\text{rel}}$
`// Step 1: Estimate individualized dose-response function using our DCNet`
Estimate $\hat{\mu}(t, x)$ via loss $\mathcal{L}_\mu$
`// Step 2: Estimate GPS using conditional normalizing flows`
Estimate $\hat{f}(t, x)$ via loss $\mathcal{L}_f$
`// Step 3: Train policy network using our reliable learning algorithm`
for $k \in \{1, 2, \ldots, K\}$ **do**
  `// K runs with random initialization`
  $\hat{\pi}_\theta^{(k)} \leftarrow$ initalize randomly
  $\lambda \leftarrow$ initalize randomly
  for *each epoch* **do**
    for *each batch* **do**
      $\mathcal{L}_\pi \leftarrow -\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}\left(\hat{\pi}_\theta^{(k)}(x_i), x_i\right) - \lambda_i\left[\hat{f}\left(\hat{\pi}_\theta^{(k)}(x_i), x_i\right) - \bar{\varepsilon}\right]\right\}$
      $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_\pi$
      $\lambda \leftarrow \lambda + \eta_\lambda \nabla_\lambda \mathcal{L}_\pi$
    end
  end
end
`// select best learned policy wrt constrained objective on validation set`
$\hat{\pi}_\theta^{\text{rel}} \leftarrow \pi_\theta^{(j)}$, with $j = \arg\max_k \sum_{i=1}^{n}\hat{\mu}\left(\pi_\theta^{(k)}(x_i), x_i\right) \cdot \mathbb{1}\left\{\hat{f}\left(\pi_\theta^{(k)}(x_i), x_i\right) \geq \bar{\varepsilon}\right\}$



Experiments

Shown: selected policy (line) and the range over 5 runs (area). The naïve baseline has a large variability across runs while ours is highly robust as expected (i.e., we see no variability).

Semi-synthetic datasets from medicine

- MIMIC-IV
  - Patients in ICU stay under mechanical ventilation
  - T = ventilation parameters
- TCGA
  - Gene expression data from cancer patients
  - T = drug combinations at chemotherapy

(a) MIMIC-IV (b) TCGA

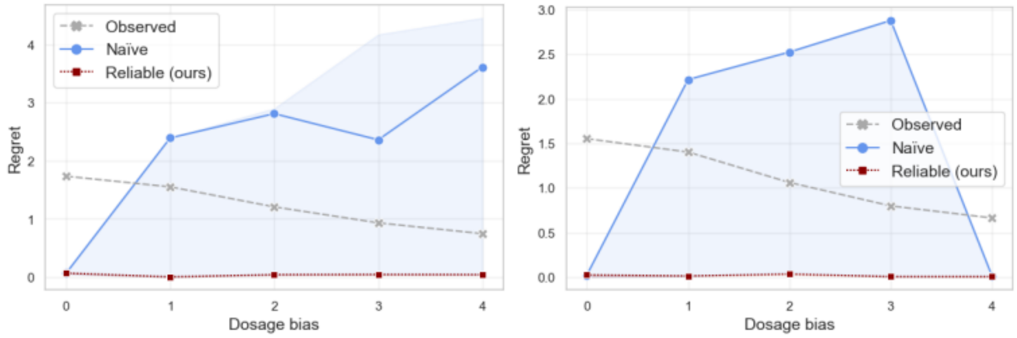


Figure 2: Robustness for dosage bias α.

(a) MIMIC-IV (b) TCGA

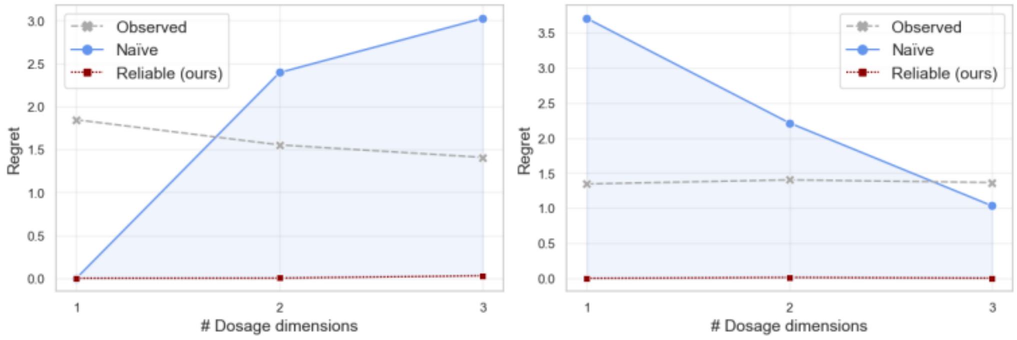


Figure 3: Robustness for number of dosages p.

Methods		MIMIC-IV			TCGA		
$\hat{\mu}$	$\hat{\pi}$	Selected	Mean	Std	Selected	Mean	Std
Oracle (μ)	observed	1.21	–	–	1.06	–	–
MLP	naïve	0.02	1.44	1.56	1.94	1.88	0.12
MLP	reliable	0.03	0.03	0.00	1.78	1.78	0.03
VCNet	naïve	2.13	2.49	0.92	3.64	2.47	1.19
VCNet	reliable	0.07	0.07	0.00	0.06	0.06	0.00
DCNet	naïve	2.81	2.07	1.17	2.52	0.94	1.21
DCNet	reliable (**ours**)	0.04	0.04	0.00	0.03	0.03	0.00

Reported: best / mean / standard deviation (lower = better)

Table 1: Performance against baselines. Regret on test set over $k = 5$ restarts.

LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

LMU MUNICH
SCHOOL OF
MANAGEMENT

INSTITUTE OF
ARTIFICIAL INTELLIGENCE
(AI) IN MANAGEMENT

Link to paper

Jonas Schweisthal
Institute of AI in Management
jonas.schweisthal@lmu.de