Policy learning under constraint: Maximization of primary outcome with control of adverse event

EUROCIM 2025

Laura FUENTES-VICENTE [1,2], Mathieu EVEN [1], Julie JOSSE [1], Antoine CHAMBAZ [2]
[1] INRIA-Idesp, Montpellier, France [2] MAP5, Paris, France

Introduction

Objective / Motivation:

- Estimate a policy maximizing a primary outcome while controlling the probability of adverse events.
- In In-Vitro-Fertilization, we aim to find the gonadotropin dose that maximizes pregnancy chances while avoiding hyperstimulation.

Background

a- Notations

- $\mathcal{O}_1, \dots, \mathcal{O}_n$, n independent **observations,** $\mathcal{O}_i = (X, A, Y, Z) \sim P_0$.
  - **Covariates** $X \in \mathcal{X}$
  - **Treatment** $A \in \{0,1\}$
  - **Primary outcome** $Y \in [0,1]$
  - **Adverse event** $\xi \in [0,1]$
- Rubin's **potential outcomes** [1] for both outcomes $(Y(0), \xi(0), Y(1), \xi(1))$.
- Complete data structures: $\mathbb{O} = (X, A, Y(1), \xi(1), Y(0), \xi(0), Y, \xi) \sim \mathbb{P}_0$.

b- Causal quantities

Conditional Average Treatment Effect (CATE) for Y

Treatment effect for Y within a subgroup $X \in \mathcal{X}$, identified under causal assumptions (SUTVA [2], positivity [3], conditional ignorability [4]),

$$(1) \quad \Delta\mu_{\mathbb{P}}(X) = E_{\mathbb{P}}[Y(1) - Y(0)|X] \triangleq \mu_{P_0}(1, X) - \mu_{P_0}(0, X) = \Delta\mu_{P_0}(X),$$

where $\mu_{P_0}(A, X) = E_{P_0}[Y|A, X]$.

Policy: maps covariates to treatment.

$$\pi: \mathcal{X} \mapsto \{0,1\}$$

Policy value: The average outcome if the policy is followed, a natural criterion to maximize, equivalent under [2,3,4] to,

$$(2) \quad \mathcal{V}_{\mathbb{P}}: \pi \mapsto E_{\mathbb{P}}[Y(\pi(X))] \approx E_P[\pi(X)\Delta\mu_P(X)].$$

Value-optimal policy:
Maximizer of (2),

$$\pi^*: x \mapsto \mathbf{1}\{\Delta\mu_{P_0}(x) > 0\}.$$

Classical policy learning

1- Direct: Predict the sign of $\Delta\mu_{P_0}(x)$.

2- Indirect: Estimation of $\Delta\mu_{P_0}(x)$.
(ex: T-learner, AIPW/DR-learner)

- x Traditional plug-in: rely on accuracy of μ_{P_0} estimates.
- x DR-learner (risk minimization): non-convex optimization, estimator $\hat{\Delta}\mu_n$ of $\Delta\mu_{P_0}$ takes problematic values.

Solution

EP-learner [5]:
Efficient plug-in risk (3) estimator.

$$(3) \quad \mathrm{R}_{P_0}: \psi \mapsto E_{P_0}[\psi(X) - 2\psi(X)\Delta\mu_{P_0}(X)]$$

1- Build efficient estimation of (3): plug-in $\Delta\mu_{P_0}$ estimator.

2- Minimize (3) for $\psi \in \Psi$.

- ✓ Stability and robustness of plug-in estimators.
- ✓ Asymptotic equivalence to oracle-efficient AIPW estimator.
- x **Overlooks adverse events.**

Contributions

1. Formulate a policy optimization problem with constraints.
2. Define a strongly convex policy in a probabilistic style (in [0, 1]).
3. Attempt to adapt the EP-learner debiasing procedure for multiple outcomes.
4. Optimize policy in a convex function space using the Frank-Wolfe algorithm.

Our problem: oracular viewpoint

We smoothly map every $\psi \in \Psi$ to a policy π, with the probability of recommending treatment (4), where,

$$(4) \quad \sigma_\beta: \Psi \to [0,1],$$
$$\pi(X) \sim Ber\left(\sigma_\beta \circ \psi(X)\right).$$

$$\Psi = Conv\{x \mapsto 2\text{expit}(X\theta) - 1: \theta \in \mathbb{R}^d\}$$

- ✓ Solves differentiability issues.
- ✓ Confidence measure in treatment recommendation.

CATE for adverse events ξ

Treatment effect of ξ within a subgroup $X \in \mathcal{X}$, identified under [2],[3],[4],

$$(5) \quad \Delta\nu_{\mathbb{P}}(X) = E_{\mathbb{P}}[\xi(1) - \xi(0)|X] \triangleq \nu_{P_0}(1, X) - \nu_{P_0}(0, X) = \Delta\nu_{P_0}(X)$$

where $\nu_{P_0}(A, X) = E_{P_0}[\xi|A, X]$, and we assume $\Delta\nu_{P_0}(X) \geq 0$.

Constraint definition:

Guarantee that policies do not increase, in average, the probability of the adverse event beyond a user-supplied $\alpha \in [0, ½]$, using $\mathbb{P}_0(\xi(0)|X)$ as baseline.

$$(6) \quad S_{P_0}: \psi \mapsto E_{P_0}[\sigma_\beta \circ \psi(X)\, \Delta\nu_{P_0}(X)] - \alpha$$

Policy optimization objective

$$(7) \quad \text{minimize } R_{P_0}(\psi) \text{ with respect to } \psi \in \Psi \text{ subject to } S_{P_0}(\psi) \leq 0$$

Problem (7) can be solved using a Lagrangian.

- ✓ Trade-off: optimizing primary outcome and controlling adverse effects.

$$(8) \quad \mathcal{L}_{P_0}: (\psi, \lambda) \mapsto R_{P_0}(\psi) + \lambda S_{P_0}(\psi)$$
$$\Psi \times \mathbb{R}^+ \to \mathbb{R}$$

Algorithm

We adapt the EP-learning algorithm.

Algorithm 1: **EP-learner for constrained policy estimation**

1- Obtain J initial cross-validated estimates of nuisance parameters: $\hat{e}_{n,j}, \hat{\mu}_{n,j}, \hat{\nu}_{n,j}$
for j = $1, \dots, J$ **do**:
  for $\beta \in B$ **do**:
    for $\lambda \in \Lambda$ **do**:
      2- Estimate the objective function in j-th fold
        (9) $\hat{\mathcal{L}}^*_{n,j}: \psi \mapsto \hat{R}^*_{n,j}(\psi) + \lambda \hat{S}^*_{n,j}(\psi)$
      3- Obtain $\hat{\psi}_{n,j}$ minimizer of (9):
        $\psi \in \text{argmin}\{\hat{\mathcal{L}}^*_{n,j}(\psi): \psi \in \Psi\}$
return $\hat{\psi}_n$

Steps 2 & 3 rely on the Frank-Wolfe algorithm.

- ✓ Avoids the need for projection.
- ✓ Convergence rate o(inverse number of iterations)

Frank-Wolfe algorithm [6]: constrained optimization algorithm expressing the optimal solution as a convex combination of extremal points of Ψ.

Ongoing work:

- We have $\hat{\mu}_{n,j}, \hat{\nu}_{n,j}$ initial cross-validated estimators of μ_0, ν_0.
- Build (9), initial plug-in estimator of (8) for all $\psi \in \Psi$,

$$(9) \quad \hat{\mathcal{L}}^{Tlr}_n(\psi) = \frac{1}{J}\sum_{j=1}^{J} \hat{\mathcal{L}}^{Tlr}_{n,j}(\psi),$$

$$\hat{\mathcal{L}}^{Tlr}_{n,j}(\psi) = E_{P^0_{n,j}}\left[\psi^2 - 2\psi\left(\hat{\mu}_{n,j}(1, X) - \hat{\mu}_{n,j}(0, X)\right) + \lambda\sigma_\beta \circ \psi(\hat{\nu}_{n,j}(1, X) - \hat{\nu}_{n,j}(0, X))\right.$$

Correct the initial estimator $\hat{\mathcal{L}}_n$ to obtain oracle efficient and convex function estimator.

Challenges:

- Non convexity of $\psi \mapsto \hat{\mathcal{L}}_n(\psi) + \mathrm{P}_n\Delta_\psi(P^0_n)$.
- Complexification of frame proposed in [5]
  - $\psi \in \Psi$ non-linear convex space
  - Correct two causal contrasts: $\Delta\mu_n$ and $\Delta\nu_n$

Ongoing results

Synthetic data:

- $n = 2\,000$
- $p = 10$
- $X \sim \text{Unif}(0,1)^{\otimes p}$
- $A \sim \text{Ber}(0.5)$
- $Y(a) = 1 - 2X_1 + X_2 - X_3 + h_Y(X, sign(a)) + \epsilon_Y$
- $\xi(a) = 1\{2 + X_1 + h_\xi(X, sign(a)) + \epsilon_\xi > 1.9\}$
- $\epsilon_Y, \epsilon_\xi \sim \mathcal{N}(0,1)$
- $h_Y(X, t) = 2(1 - X_1 - X_2)t$
- $h_\xi(X, t) = (1 + X_1 - X_2)t$

Optimal solutions for fixed λ:

- $\psi_{\lambda,0} \in \text{argmin}\{\mathcal{L}_{P_0,\lambda}(\psi): \psi \in \Psi\}$
- $\hat{\psi}_{\lambda,n} \in \text{argmin}\{\hat{\mathcal{L}}^{Tlr}_{n,\lambda}(\psi): \psi \in \Psi\}$

Figure 1: Synthetic setting.

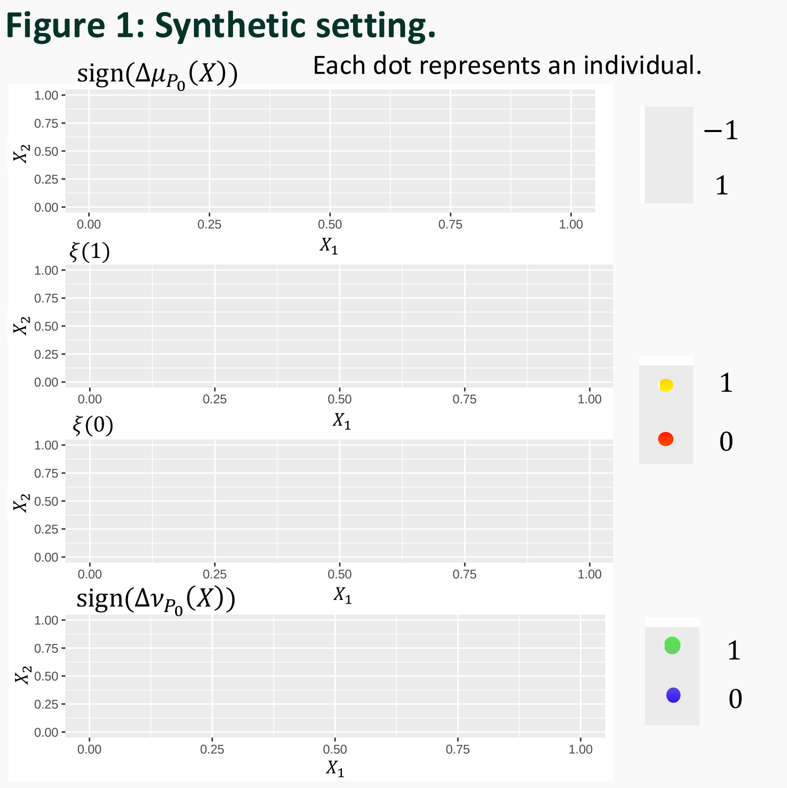


Figure 2: Evaluation of optimal solutions $\psi_{\lambda,0}$ (left), $\hat{\psi}_{\lambda,n}$ (right) with respect to λ

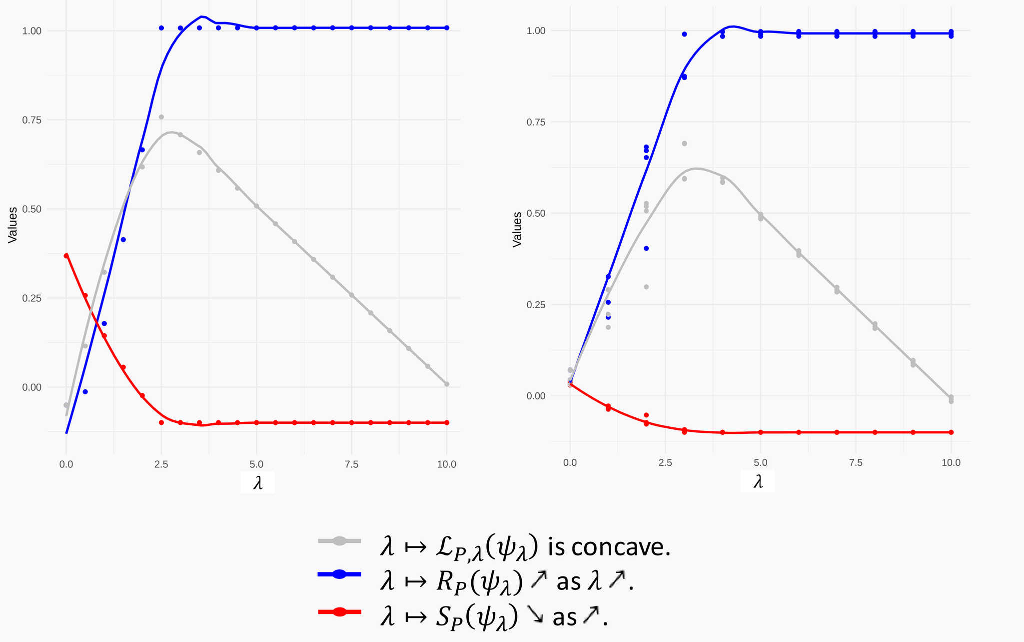
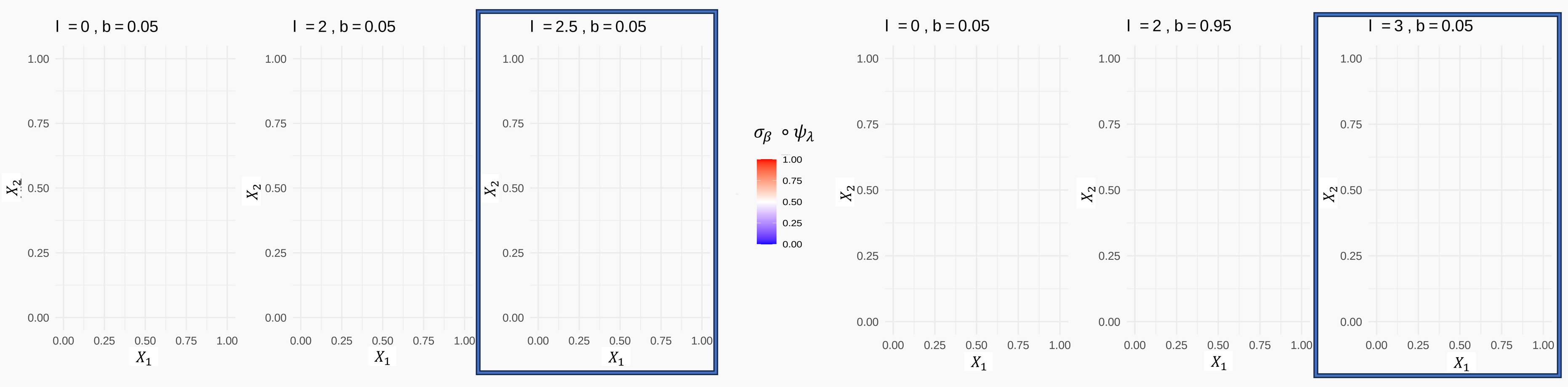


Figure 2: Treatment recommendations $\sigma_\beta \circ \psi_{\lambda,0}$ (left) $\sigma_\beta \circ \hat{\psi}_{\lambda,n}$ (right) for different λ



- $\lambda \nearrow \sigma_\beta \circ \psi_\lambda \to 0$ for all individuals.
- Estimated policy is similar to oracular policy

Perspectives

- Further develop the first EP-learning step, for estimation of (8).
- Speed-up Frank-Wolfe algorithm.
- Compare to state of the art constrained policy optimization techniques.

References

[1] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100(469):322–331
[2] Imbens, G. W. and Rubin, D. B. (2015).Causal inference in statistics, social, and biomedical sciences. Cambridge university press
[3] Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment Journal of the American statistical association, 75(371):591–593
[4] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects Biometrika, 70(1):41
[5] van der Laan, L., Carone, M., and Luedtke, A. (2024). Combining t-learning and dr-learning: a framework for oracle-efficient estimation of causal contrasts arXiv preprintarXiv:2402.01972
[6] Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In International conference on machine learning, pages 427–435. PMLR.