

Doubly Robust Fusion of Many Treatments for Policy Learning

Ke Zhu

Department of Statistics, North Carolina State University
Department of Biostatistics and Bioinformatics, Duke University

Joint work with Jianing Chu, Ilya Lipkovich , Wenyu Ye, and Shu Yang

June 13–19, 2025

Forty-Second International Conference on Machine Learning (ICML), Vancouver

Motivation

Precision medicine: Tailors treatment to patient characteristics to account for treatment effect heterogeneity and improve patient outcomes.

Goal: Learn **individualized treatment rules** (ITRs, or policy) from observational data.

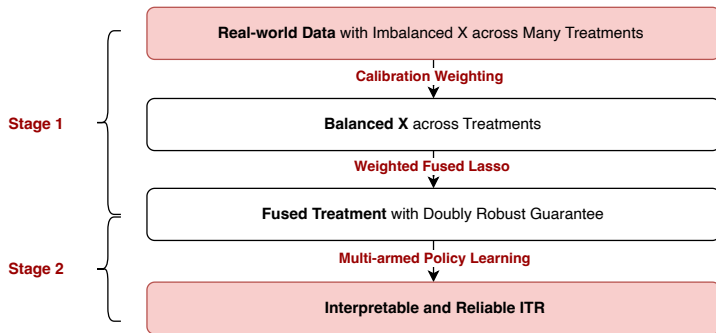
Challenges:

- **Many treatment arms**, but limited sample per arm.
- **Covariate imbalance** across treatment groups.

Limitations of existing methods:

- *Multi-armed policy learning* fails with many treatments (Zhou, Athey, and Wager, 2023).
- *Linear fusion without covariate balancing* is sensitive to misspecification (Ma, Zeng, and Liu, 2022).

Solution: Two-Stage Framework



Double robustness

Fused ITR is valid if either the *outcome* or *weighting* model is correct.

Problem Setup

Potential outcome: $Y(a)$, where $a \in \mathcal{A} \equiv \{1, \dots, K\}$.

Data: p -dimensional covariates $X_i \in \mathcal{X}$, treatment A_i , outcome Y_i .

Goal: Learn an ITR $d : \mathcal{X} \rightarrow \mathcal{A}$ that maximizes the value $\mathbb{E}[Y(d(X))]$.

Assumptions:

- Consistency: $Y = Y(A)$.
- Unconfoundedness: $(Y(1), \dots, Y(K)) \perp\!\!\!\perp A \mid X$.
- Positivity: $\mathbb{P}(A = a \mid X = x) > \pi_{\min} > 0$.

Propensity score: $\pi_a(x) \equiv \mathbb{P}(A = a \mid X = x)$.

Conditional outcome mean: $\mu_a(x) \equiv \mathbb{E}[Y(a) \mid X = x]$.

Oracle Group Structure

Among K treatments, there are M latent groups $\mathcal{G}_1^*, \dots, \mathcal{G}_M^*$:

- **Within-group:** $\mu_a(x) = \mu_{a'}(x)$ for $a, a' \in \mathcal{G}_b^*$.
- **Between-group:** $\mu_a(x) \neq \mu_{a'}(x)$ for $a \in \mathcal{G}_b^*, a' \in \mathcal{G}_{b'}^*, b \neq b'$.

Exact equality ensures identifiability and theoretical guarantees.

Fused Lasso allows approximate grouping in practice.

Stage 1: Calibration-Weighted Treatment Fusion

Step 1: Calibration Weighting

For each treatment $a \in \mathcal{A}$, solve for **weights** $\{\hat{w}_i : A_i = a\}$:

$$\begin{aligned} \min_{w_i: A_i=a} \sum_{i: A_i=a} h_\gamma(w_i), \quad & \text{(minimize deviation from uniform weights)} \\ \text{s.t.} \quad \sum_{i: A_i=a} w_i X_i &= \bar{X}, \quad \text{(covariate balance)} \quad \sum_{i: A_i=a} w_i = 1. \quad \text{(normalization)} \end{aligned}$$

where $h_\gamma(w) = \frac{(n_a w)^{\gamma+1} - 1}{\gamma(\gamma+1)}$ is from the **Cressie-Read family**.

Special cases:

- $\gamma = 0$ gives **entropy balancing** ($\sum w_i \log w_i$).
- $\gamma = -1$ gives **empirical likelihood** ($\sum \log w_i$).

Stage 1: Calibration-Weighted Treatment Fusion

Step 2: Weighted Fused Lasso

Fit a weighted working model with pairwise fusion:

$$\hat{\zeta} = \min_{\zeta} \left\{ \frac{1}{n} \sum_{a \in \mathcal{A}} \sum_{i: A_i = a} \hat{w}_i \mathcal{L}(Y_i - M_0(X_i), X_i^\top \zeta_a) + \sum_{1 \leq a < a' \leq K} p_{\lambda_n}(\|\zeta_a - \zeta_{a'}\|_1) \right\}.$$

- \mathcal{L} is a loss function (e.g., squared error for continuous outcomes).
- $M_0(X)$ is a nuisance main effect estimated separately.
- p_{λ_n} is a fusion penalty.

Stage 1: Calibration-Weighted Treatment Fusion

Consistency of oracle estimator $\hat{\zeta}^{\text{or}}$: Under regularity and *double robustness*, with known latent group structure,

$$\|\hat{\zeta}^{\text{or}} - \zeta^*\|_{\infty} \leq C\sqrt{p n \log(n)}/N_{\min},$$

where $N_{\min} = \min_{b \in \mathcal{B}} \sum_{i=1}^n \mathbb{I}\{A_i \in \mathcal{G}_b^*\}$ is the smallest group size.

Oracle property of $\hat{\zeta}$: If the between-group signal is strong and the penalty is properly tuned,

$$\mathbb{P}(\hat{\zeta} = \hat{\zeta}^{\text{or}}) \rightarrow 1.$$

Implication: Under a completeness condition, equal $\hat{\zeta}_a$'s recover the oracle groups $\{\mathcal{G}_b^*\}$.

Stage 2: Multi-Armed Policy Learning

Step 1: Estimate $\mu_b(x)$ and $\pi_b(x)$ for fused group b .

Step 2: Evaluate each policy $d^{\mathcal{B}}(x)$ with cross-fitted AIPW:

$$\hat{V}(d^{\mathcal{B}}) = \frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{I}\{B_i = d^{\mathcal{B}}(X_i)\} \frac{Y_i - \hat{\mu}_{B_i}^{-l(i)}(X_i)}{\hat{\pi}_{B_i}^{-l(i)}(X_i)} + \hat{\mu}_{d^{\mathcal{B}}(X_i)}^{-l(i)}(X_i) \right\}.$$

Step 3: Optimize $d^{\mathcal{B}}$ over policy class $\mathcal{D}^{\mathcal{B}}$ (e.g., depth- D trees).

Stage 2: Multi-Armed Policy Learning

General regret bound: Under regularity conditions and *rate doubly robust model assumption* (i.e., at least one nuisance model is consistent and their product error is $o(n^{-1})$),

$$R(\hat{d}^{\mathcal{B}}) = O_{\mathbb{P}} \left(\kappa(\mathcal{D}^{\mathcal{B}}) \sqrt{V_*/n} \right),$$

where $\kappa(\mathcal{D}^{\mathcal{B}})$ quantifies policy class complexity and V_* is the worst-case variance.

Policy tree regret bound: For depth- D policy trees,

$$R(\hat{d}^{\mathcal{B}}) = O_{\mathbb{P}} \left(\left\{ \sqrt{(2^D - 1) \log p + 2^D \log M} + \frac{4}{3} D^{1/4} \sqrt{2^D - 1} \right\} \sqrt{V_*/n} \right).$$

Simulation Setup

Data

- $K = 16$ treatments, $M = 4$ latent groups
- Covariate shift and sample size imbalance

Competing Methods

- **Policy tree** without fusion
- **Fusion + policy tree** without calibration weighting (CW)
- **Ma, Zeng, and Liu (2022)**: linear fusion + ITR without weighting

Metrics

- **ARI**: measures fusion quality (1 = perfect)
- **Number of groups**: oracle = 4
- **Value**: $\mathbb{E}[Y(d(X))]$ (higher is better)
- Monte Carlo standard errors in parentheses (200 runs)

Simulation Results: Misspecified Outcome Model

Nonlinear $\mu_a(X)$; all X used for weighting

Method	ARI \uparrow	# Groups	Value \uparrow
Policy tree (baseline)	–	16.000	8.77 (0.08)
Fusion + policy tree	0.26 (0.14)	10.73 (1.93)	8.78 (0.09)
CW + fusion + policy tree (proposed)	0.96 (0.06)	4.34 (0.60)	8.89 (0.11)
Ma, Zeng, and Liu (2022)	0.26 (0.14)	10.73 (1.93)	8.51 (0.12)

Simulation Results: Misspecified Weighting Model

Linear $\mu_a(X)$; partial X used for weighting

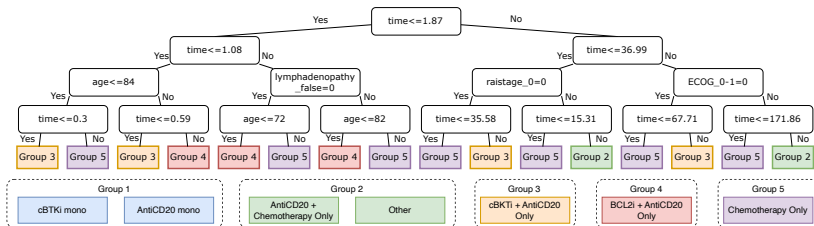
Method	ARI \uparrow	# Groups	Value \uparrow
Policy tree (baseline)	–	16.00	6.35 (0.06)
Fusion + policy tree	0.88 (0.13)	5.42 (1.42)	6.41 (0.04)
CW + fusion + policy tree (proposed)	0.96 (0.06)	4.46 (0.66)	6.43 (0.02)
Ma, Zeng, and Liu (2022)	0.88 (0.13)	5.42 (1.42)	6.39 (0.00)

- Chronic Lymphocytic Leukemia (CLL) and Small Lymphocytic Lymphoma (SLL) are slow-progressing blood cancers with **complex treatment options**.
- **Outcome:** overall survival (binary).
- **Covariates (10):** race, region, PayerBin, SES Index, gender, ECOG score, Rai stage, lymphadenopathy, age at LOT start, time from diagnosis to LOT.
- **ITR learning:** excluded race, region, SES proxies (PayerBin, SES Index) for fairness.

Real Data Application: CLL/SLL Patients

Treatment	Number of Patients
cBTKi mono	3392
AntiCD20 + Chemotherapy Only	1726
AntiCD20 mono	1230
BCL2i + AntiCD20 Only	463
cBTKi + AntiCD20 Only	408
Chemotherapy Only	215
Other	412
Total	10346

Real Data Application: CLL/SLL Patients



- Group 1 includes *two monotherapies* with similar mechanisms and intensity.
- *Combination therapies* and *chemotherapy-only* form distinct treatment groups.
- **Older or recently diagnosed patients** tend to be assigned to chemotherapy-only.
- **Younger or long-diagnosed patients** are guided to combination therapies.

Takeaway Messages



Challenge: Learn interpretable and reliable ITRs from observational data with many treatments, limited samples per arm, and covariate imbalance.

Solution: A novel two-stage framework that integrates calibration weighting, fused lasso, and interpretable policy learning.

Guarantees: Doubly robust theory for both stages, ensuring oracle recovery and providing regret bounds, supported by strong empirical results in simulations and real-world data.

Thank you!

References

-  Ma, Haixu, Donglin Zeng, and Yufeng Liu (2022). “Learning individualized treatment rules with many treatments: A supervised clustering approach using adaptive fusion”. In: *Advances in Neural Information Processing Systems* 35, pp. 15956–15969.
-  Zhou, Zhengyuan, Susan Athey, and Stefan Wager (2023). “Offline multi-action policy learning: Generalization and optimization”. In: *Operations Research* 71.1, pp. 148–183.