

Fairness with Constrained Paths

for the seminar: *Fairness in Machine Learning*,
organized by M. Hardt, Fall 2017, UC Berkeley,
given Tuesday 12th December, 2017

Nima Hejazi

Group in Biostatistics
University of California, Berkeley
stat.berkeley.edu/~nhejazi



nimahejazi.org
[twitter/@onshejazi](https://twitter.com/onshejazi)
[github/nhejazi](https://github.com/nhejazi)

[slides: goo.gl/qc7JPH](https://goo.gl/qc7JPH)



This slide deck is for a short presentation on the final project undertaken for “Fairness in Machine Learning,” a special topics seminar organized by Moritz Hardt in the Fall 2017 term at UC Berkeley.

Slides: <https://goo.gl/xZ5qPa>

With notes: <https://goo.gl/qc7JPH>

Preview: Summary

- ▶ Targeted Learning provides a framework for estimating complex parameters in nonparametric (infinite-dimensional) statistical models.
- ▶ Constraints may be imposed as functionals defined over the target parameter of interest.
- ▶ Estimating constrained parameters may be seen as iteratively minimizing a loss function along a *constrained* path in the parameter space Ψ .
- ▶ Optimal nonparametric estimates of parameters (and constituent parts thereof) may be obtained with Super Learning (i.e., stacked regression, ensemble models).

1

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

Why Nonparametrics?

Everyone believes in the normal law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.

-Henri Poincaré

Obviously, it's important to explain the motivating example here.

Background, Data, Notation

- ▶ Data: $O = (W, X, Y)$, where W is baseline covariates (e.g., sex, height), X is a sensitive characteristic, Y an outcome of interest.
- ▶ Consider n i.i.d. copies O_1, \dots, O_n of $O \sim P_0 \in \mathcal{M}$.
- ▶ Here, \mathcal{M} is an infinite-dimensional statistical model (i.e., indexed by an infinite-dimensional vector).
- ▶ We discuss the estimation of a target parameter $\psi : \mathcal{M} \rightarrow \psi(\mathcal{M})$.

▶

$$\Psi(P) = \arg \min_{\psi \in \Psi} \mathbb{E}_P L(\psi)$$

This is just the setup and all stuff that's been done already. We just need to see this to get a feel for what's going to be happening with the derivation of constraint-specific paths.

Constrained Functional Parameters I

- ▶ Consider estimating $\Psi(P) = \mathbb{E}_P(Y \mid W, X)$ from $O = (W, X, Y) \sim \mathcal{M}$, with loss $L_Y(\psi) = -(Y \cdot \log(P(Y \mid X, W)) + (1 - Y) \cdot \log(1 - P(Y \mid X, W)))$.
- ▶ Let $\Theta_\psi(P) : \mathcal{M} \rightarrow \mathbb{R}$ be a parameter for each $\psi \in \Psi$. Further, let $\Theta_\psi(P)$ be pathwise differentiable.
- ▶ e.g., *equalized odds*: $\Theta_\psi(P_0) = \sum_y \{ \mathbb{E}_{P_0}(L(\psi)(O) \mid X = 1, Y = y) - \mathbb{E}_{P_0}(L(\psi)(O) \mid X = 0, Y = y) \}^2$

▶

$$\Psi(P) = \arg \min_{\psi \in \Psi, \Theta_\psi(P)=0} \mathbb{E}_P L(\psi)$$

Constrained Functional Parameters II

- We wish to estimate $\Psi^*(P)$, the projection of $\Psi(P)$ onto the subspace $\Psi^*(P) = \{\psi \in \Psi : \Theta_\psi(P) = 0\}$ w.r.t. loss-based dissimilarity:

$$d_P(\psi, \Psi(P)) = \mathbb{E}_P L(\psi) - \mathbb{E}_P L(\Psi(P)).$$

►

$$\Psi^*(P) = \arg \min_{\psi \in \Psi, \Theta_\psi(P)=0} d_P(\psi, \Psi(P))$$

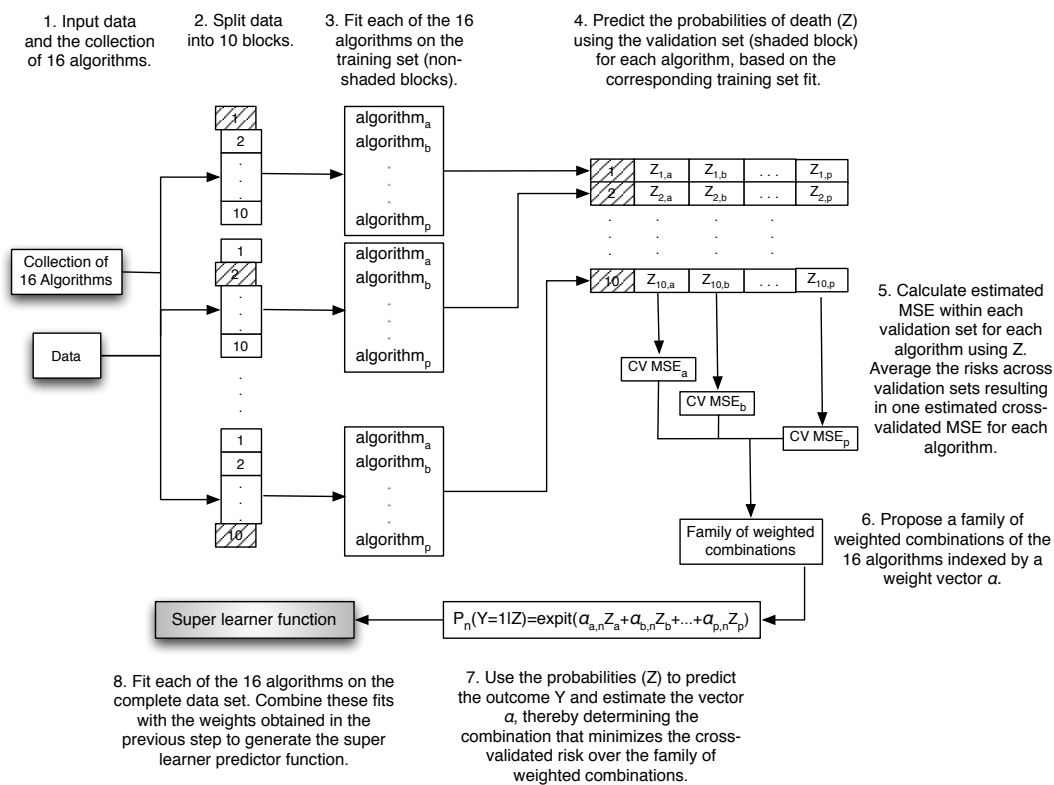
$$(\Psi^*, \lambda) = (\Psi^*(P), \Lambda(P)) \equiv \arg \min_{\psi \in \Psi, \lambda} \mathbb{E}_P L(\psi) + \lambda \Theta_\psi(P), \quad (1)$$

- *Lemma:* If $\tilde{\Psi}(P) = (\Psi^*(P), \Lambda(P))$ is the minimizer of the Lagrange multiplier penalized loss (??), then it readily follows that $\Psi^*(P) = \arg \min_{\psi \in \Psi, \Theta_\psi(P)=0} \mathbb{E}_P L(\psi)$.

Intuition for Super Learner I

- ▶ For a random variable $O = (W, X, Y)$, let the oracle selector be a rule that picks the algorithm with the lowest cross-validated risk under the true probability distribution P_0 .
- ▶ Asymptotic results prove that in realistic scenarios (where none of the algorithms represent the true relationship), the “discrete Super Learner” performs asymptotically as well as the oracle selector (given the algorithms in the collection).

Intuition for Super Learner II



Super Learner for Constrained Parameters I

- Consider the risk function for $\tilde{\psi} = (\psi^*, \lambda)$ — i.e.,
 $R(\tilde{\psi} \mid P) \equiv PL(\psi^*) + \lambda\Theta(\psi^* \mid P)$.

►

$$\mathbb{E}(\Psi^*(P), \Lambda(P)) = \arg \min_{(\psi, \lambda) \in \Psi \times \mathbb{R}_{\geq 0}} R(\psi, \lambda \mid P)$$

- For a given estimator $\tilde{\psi}(P_n) = (\hat{\Psi}^*(P_n), \hat{\lambda}(P_n))$ of $\tilde{\Psi}(P_0)$, and a cross-validation scheme defined by a distribution of $B_n \in \{0, 1\}^n$ that splits the sample.

- We can define a conditional risk as follows:

$$R_0(\tilde{\psi}, P_n) = \mathbb{E}_{B_n} P_0 L(\hat{\Psi}^*(P_{n, B_n}^0)) + \hat{\lambda}(P_n) \mathbb{E}_{B_n} \Theta(\hat{\Psi}^*(P_{n, B_n}^0) \mid P_0)$$

8

- Here P_{n, B_n}^0 denotes the empirical distribution of the training sample.

Super Learner for Constrained Parameters II

$$R_{n,CV}(\tilde{\psi}, P_n) = E_{B_n} P_{n,B_n}^1 L(\hat{\Psi}^*(P_{n,B_n}^0)) + \hat{\Lambda}(P_n) E_{B_n} \Theta(\hat{\Psi}^*(P_{n,B_n}^0) \mid P_{n,B_n}^*) \quad (2)$$

- ▶ Given a set of candidate estimators $\tilde{\psi}_j(P_n) = (\hat{\Psi}_j^*(P_n), \hat{\Lambda}_j(P_n))$, $j = 1, \dots, J$, the cross-validation selector is given by:
 $J_n = \arg \min_j R_{n,CV}(\tilde{\psi}_j, P_n)$
- ▶ A Super Learner may be defined over a set of candidate nonparametric regression functions (i.e., machine learning algorithms).
- ▶ Here, we may define a Super Learner of $\tilde{\Psi}$ by
 $\tilde{\psi}_n \equiv \tilde{\psi}_{J_n}(P_n) = (\hat{\Psi}_{J_n}(P_n), \hat{\lambda}_{J_n}(P_n))$

Mappings with Constrained Super Learners

A straightforward approach to generating estimators of the constrained parameter would be to simply generate a mapping according to the following simple process:

1. Generate an unconstrained Super Learner ψ_n of the unconstrained parameter ψ_0 ,
2. Map an estimator $\Theta_{\psi_n, n}$ of the constraint $\Theta_{\psi_n}(P_0)$ into the path $\psi_{n, \lambda}$ mentioned above. The corresponding solution $\psi_n^* = \psi_{n, \lambda_n}$ of $\Theta_{\psi_n, \lambda_n, n} = 0$ generates an estimator of the constrained parameter.

Constraint-specific Paths I

Let $P_0 \in \mathcal{M}$ be the true distribution of the data $O = (W, X, Y)$, contained in an infinite-dimensional statistical model \mathcal{M} .

Consider the solution

$\psi_{0,\lambda} = \arg \max_{\psi \in \Psi} \mathbb{E}_{P_0} L(\psi) + \lambda \Theta_0(\psi)$, noting that $\{\psi_{0,\lambda} : \lambda\}$ represents a path in the parameter space Ψ through ψ_0 at $\lambda = 0$, which we refer to as the *constraint-specific path*.

We then leverage this construction to map an initial estimator of the unconstrained parameter ψ_0 into its corresponding constrained version ψ_0^* .

Constraint-specific Paths II

The fine print:

Theorem

Let $\{\psi_{\delta,H} : \delta\}$ be a one-dimensional path through ψ at $\delta = 0$ with direction $H \in \mathbf{H}(\psi)$, where H varies over a set of directions \mathcal{H} . Let $H(\psi)$ be a Hilbert space with inner product $\langle f, g \rangle$.

Let $T(\psi) \subset H(\psi)$ be the closure of the linear span of \mathcal{H} . Let $D_{0,L}(\psi) \in T(\psi)$ be the canonical gradient of the pathwise derivative $\left. \frac{d}{d\delta} \mathbb{E}_{P_0} L(\psi_{\delta,H}) \right|_{\delta=0} = \langle D_{0,L}(\psi), H \rangle$. Let $D_{0,\Theta}(\psi) \in T(\psi)$ be the canonical gradient of the pathwise derivative $\left. \frac{d}{d\delta} \Theta_0(\psi_{\delta,H}) \right|_{\delta=0} = \langle D_{0,\Theta}(\psi), H \rangle$. Let $\psi_{0,\lambda} = \arg \min_{\psi \in \mathcal{H}} \mathbb{E}_{P_0} L(\psi) + \lambda \Theta_0(\psi)$.

Suppose that $\psi_{0,\lambda}$ solves its score equations $\left. \frac{d}{d\delta} \mathbb{E}_{P_0} L(\psi_{0,\lambda,\delta,H}) \right|_{\delta=0}$ for all paths. Then, we have

$$0 = D_L(\psi_{0,\lambda}) + \lambda D_{0,\Theta}(\psi_{0,\lambda}). \quad (3)$$

...

This now uniquely defines the solution $\psi_{0,\lambda}$ and an algorithm for computing it, given above. Finally, letting $\psi_0^* = \arg \min_{\psi \in \mathcal{H}, \Theta_0(\psi)=0} \mathbb{E}_{P_0} L(\psi)$, we have $\psi_0^* = \psi_{0,\lambda_0}$, where λ_0 is such that $\Theta_0(\psi_{0,\lambda}) = 0$.

Future Work

- ▶ Further generalization of constraint-specific paths: the solution path $\{\psi_{0,\lambda} : \lambda\}$ in the parameter space Ψ through ψ_0 at $\lambda = 0$.
- ▶ Further develop relation between constraint-specific paths and universal least favorable submodels (Efron's least favorable families).
- ▶ Integration of the approach of constraint-specific paths with classical targeted maximum likelihood estimation — in particular, what, if any, are the implications for inference?
- ▶ Try our approach out with practical constraint-type problems (e.g., fairness via equalized odds, physician knowledge in prediction).

Review: Summary

- ▶ Targeted Learning provides a framework for estimating complex parameters in nonparametric (infinite-dimensional) statistical models.
- ▶ Constraints may be imposed as functionals defined over the target parameter of interest.
- ▶ Estimating constrained parameters may be seen as iteratively minimizing a loss function along a *constrained* path in the parameter space Ψ .
- ▶ Optimal nonparametric estimates of parameters (and constituent parts thereof) may be obtained with Super Learning (i.e., stacked regression, ensemble models).

14

It's always good to include a summary.

References I

15

Acknowledgments

Mark J. van der Laan

University of California, Berkeley

Funding source:

National Library of Medicine (of NIH): T32LM012417

16

Thank you.

Slides: goo.gl/xZ5qPa



Notes: goo.gl/qc7JPH

stat.berkeley.edu/~nhejazi

nimahejazi.org

[twitter/@nshejazi](https://twitter.com/nshejazi)

[github/nhejazi](https://github.com/nhejazi)

Here's where you can find me, as well as the slides for this talk.