Efficient Policy Learning

Stefan Wager
Stanford University

Innovative Statistics and
Machine Learning for Precision Medicine
University of Minnesota, 15 September 2017

Joint work with Susan Athey

Stanford Hospital and Clinics: Discharge criteria for post-anesthesia care.

- Consciousness score: ≥ 1 out of 2.
- Respiration score: 2 out of 2.
- Blood pressure score: ≥ 1 out of 2.
- ...

Total score must be ≥ 10 out of 12.

Statistical Setup

We want to **learn a policy** π that can be applied in the future:

$$\pi : \mathcal{X} \to \{\pm 1\}, \quad \pi \in \Pi.$$

To do so, we have access to **observational data** collected in the past. In order to predict the effect of policy changes, we need to identify and estimate the **causal effect** of the treatment.

- We have **i.i.d. observations** $(X_i, Y_i, W_i) \in \mathcal{X} \times \mathbb{R} \times \{\pm 1\}$ for $i = 1, ..., n$, where W_i is the treatment assignment.
- Following the Neyman-Rubin model, we posit **potential outcomes** $\{Y_i(\pm 1)\}$ corresponding to how i-th subject would have responded to different W_i, such that $Y_i = Y_i(W_i)$.
- To identify treatment effects, we assume **unconfoundedness** (Rosenbaum and Rubin, 1983), $\{Y_i(-1), Y_i(+1)\} \perp\!\!\!\perp W_i \,\big|\, X_i$, and **overlap**.

What is Policy Learning?

We interpret Y_i as a utility, and specify the **optimal policy** $\pi^* := \text{argmax}\{\mathbb{E}\left[Y(\pi(X))\right] : \pi \in \Pi\}$ as the one that maximizes **expected utility**. Equivalently

$$\pi^* := \text{argmax}\left\{Q(\pi) : \pi \in \Pi\right\},$$
$$Q(\pi) := 2\mathbb{E}\left[Y(\pi(X)) - \frac{Y(-1) + Y(+1)}{2}\right],$$

where $Q(\pi)$ measures the improvement over a randomized baseline.

As in Manski (2004) or Qian and Murphy (2011), we control **minimax regret** (Savage, 1951). We define regret as $R(\pi)$, and want to learn a policy $\hat{\pi}$ with **uniform bounds** on regret:

$$R\left(\hat{\pi}\right) \ll 1, \text{ where } R(\pi) := \frac{Q\left(\pi^*\right) - Q(\pi)}{2}.$$

Minimax Regret Policy Learning

If the class of candidate policies Π has **no structure**, e.g., $\mathcal{X}$ is discrete, Π contains all $2^{|\mathcal{X}|}$ possible treatment assignments, the problem of minimax regret policy learning is well understood.

Results by Manski (2004, 2009), Hirano and Porter (2009) and Stoye (2009) all confirm that minimax policy learning reduces to efficient **treatment effect estimation**. The asymptotically minimax choice $\hat{\pi}^*(\cdot)$ is of the form

$$\hat{\pi}^*(x) = \mathbf{1}\left(\{\hat{\tau}(x) > c(x)\}\right),$$

where $\hat{\tau}(x)$ is an efficient estimate of the conditional ATE,

$$\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X = x\right].$$

A limitation of these results, however, is that they give the practitioner no control over the **form of the policy** $\hat{\pi}(\cdot)$.

Statistical Setup, Revisited

We posit **unconfounded** observations via **potential outcomes** $(X_i,\ Y_i(-1),\ Y_i(+1),\ W_i)$, with $Y_i = Y_i(W_i)$, and write

$$\mu_w(x) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right], \quad e_w(x) = \mathbb{P}\left[W_i = w \,\middle|\, X_i = x\right].$$

Throughout, we will assume that $\mu_w(\cdot)$ and $e(\cdot)$ belong to a **non-parametric class** that allows for $o(n^{-1/4})$-consistent estimation under L_2 error; use **cross-fitting** for regularity (Chernozhukov et al., 2017).

We want to learn a **policy** $\pi : \mathcal{X} \to \{\pm 1\}$ such that $\pi \in \Pi$, where Π is a "simple" class of functions. We will assume that Π has a **finite VC-dimension** or, more generally, a finite entropy integral.

Statistical Setup, Revisited

Considering **different function classes** for $\mu_w(\cdot)$ and $e(\cdot)$ versus $\pi(\cdot)$ may appear strange, but is essential in many applications.

The functions $\mu_w(\cdot)$ and $e(\cdot)$ need to **describe nature**. Using more pre-treatment features (usually) helps unconfoundedness,

$$\{Y_i(-1),\ Y_i(+1)\} \perp\!\!\!\perp W_i \,\big|\, X_i.$$

Conversely, the policy $\pi(\cdot)$ must be **implementable in practice**. Features we can use for $\mu_w(\cdot)$ and $e(\cdot)$ but not $\pi(\cdot)$ include:

- **Unreliably available features** (e.g., collected by specialist).
- **Gameable features** (e.g., self-reported preferences).
- **Legally protected classes** (e.g., religion, national origin).

Moreover, we may want Π to encode constraints on:

- **Total budget** or marginal **subgroup treatment rates**.
- **Functional form** for easier implementation or audit.

We study policy learning in a way that is aware of such constraints.

A First Solution

A natural approach is to optimize an **estimated value function**,

$$\hat{\pi} = \text{argmax} \left\{ \widehat{Q}(\pi) : \pi \in \Pi \right\}.$$

A simple, **unbiased estimate** of $Q(\pi)$ is (remarkably?) available:

$$\begin{aligned} &\mathbb{E}\left[\pi(X_i) W_i Y_i \,/\, \mathbb{P}\left[W = W_i \,\middle|\, X = X_i\right]\right] \\ &\quad = \mathbb{E}\left[Y(\pi(X_i))\right] - \mathbb{E}\left[Y(-\pi(X_i))\right] \\ &\quad = 2\mathbb{E}\left[Y(\pi(X_i)) - \left(Y_i(+1) - Y_i(-1)\right)/2\right] = Q(\pi). \end{aligned}$$

This insight, along with the induced policy learner (a.k.a **outcome weighted learning**),

$$\hat{\pi}_{IPW} = \text{argmax} \left\{ \sum_{i=1}^{n} \frac{\pi(X_i) W_i Y_i}{\mathbb{P}\left[W = W_i \,\middle|\, X = X_i\right]} : \pi \in \Pi \right\},$$

has been independently studied across several fields, including **statistics** (Zhao, Zeng, Rush and Kosorok, 2012), **machine learning** (Dudík et al., 2011; Swaminathan and Joachims, 2015), and **economics** (Kitagawa and Tetenov, 2015).

Inverse-Propensity Policy Learning

The inverse-propensity weighted method uses

$$\hat{\pi}_{IPW} = \text{argmax} \left\{ \sum_{i=1}^{n} \frac{\pi(X_i) W_i Y_i}{\mathbb{P}\left[W = W_i \,\middle|\, X = X_i\right]} : \pi \in \Pi \right\}.$$

The resulting procedure is consistent, with **policy regret bounds** (Kitagawa and Tetenov, 2015; Swaminathan and Joachims, 2015):

$$Q(\pi^*) - Q(\hat{\pi}_{IPW}) = \mathcal{O}_P \left(\frac{\sup\{|Y|\}}{\inf\left\{\mathbb{P}\left[W = w \,\middle|\, X\right]\right\}} \sqrt{\frac{VC(\Pi)}{n}} \right).$$

However, the estimator is not **translation invariant** in Y_i (and neither is the regret bound).

There are several proposals for improvement, including Dudík et al. (2011), Zhang et al. (2012) and Zhou et al. (2015); however, **existing theory gives no guidance** on which method to prefer.

Efficient Policy Evaluation

We start by considering policy **evaluation**. Inverse-propensity weighted policy learning optimizes

$$\widehat{Q}_{IPW}(\pi) := \sum_{i=1}^{n} \frac{\pi(X_i) W_i Y_i}{\hat{e}_w(X_i)}.$$

We can interpret $\widehat{Q}(\pi)$ as an **average treatment effect** estimate,

$$Q(\pi) = \left(\mathbb{E}\left[Y(\pi(X))\right] - \mathbb{E}\left[Y(-\pi(X))\right]\right),$$

where "treated" people get policy $\pi(\cdot)$ and controls get $-\pi(\cdot)$.

Efficient estimation of $Q(\pi)$ is well understood (Bickel et al., 1998; Hahn, 1998; Hirano et al., 2003; Robins & Rotnitzky, 1995):

$$\widehat{Q}_{DR}(\pi) = \sum_{i=1}^{n} \pi(X_i) \left(\hat{\mu}_+(X_i) - \hat{\mu}_-(X_i) + W_i \frac{Y_i - \hat{\mu}_{W_i}(X_i)}{\hat{e}_{W_i}(X_i)}\right).$$

Can we leverage this insight for policy **learning**?

Efficient Treatment Effect Estimation

We consider $\hat{\pi}_{DR}$ be the maximizer of $\widehat{Q}_{DR}$ over Π, where

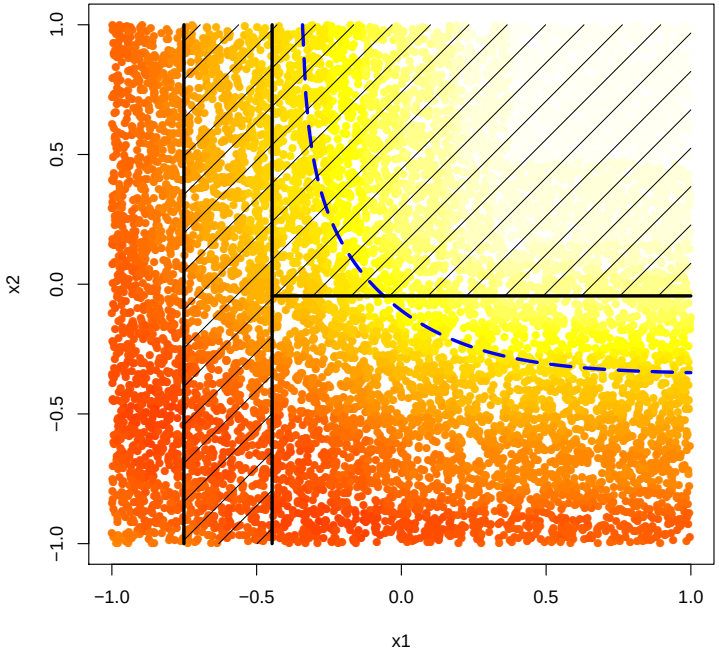
$$\widehat{Q}_{DR}(\pi) = \sum_{i=1}^{n} \pi(X_i) \left(\hat{\mu}_+(X_i) - \hat{\mu}_-(X_i) + W_i \frac{Y_i - \hat{\mu}_{W_i}(X_i)}{\hat{e}_{W_i}(X_i)} \right).$$

It is well known that $\widehat{Q}_{DR}$ has the **efficient asymptotic variance** for evaluating any single policy π.

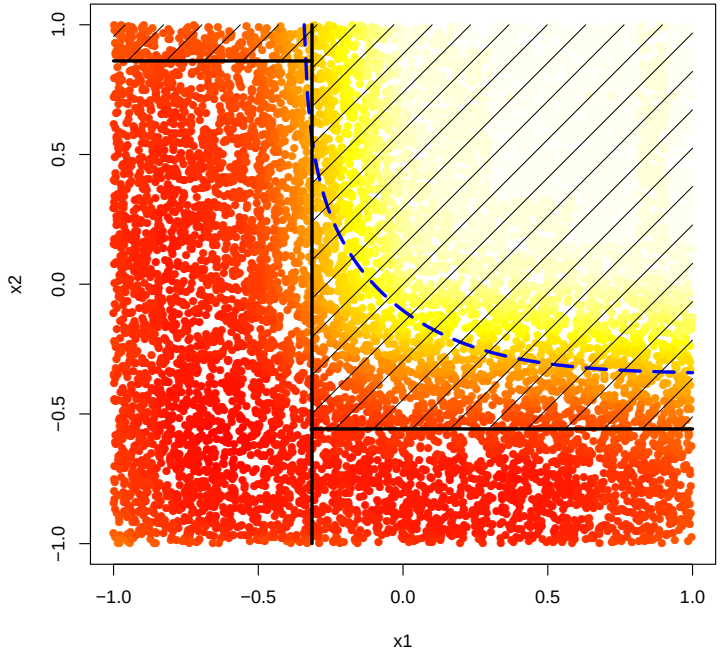
Question: Can we translate efficiency results about evaluating single policies into results about policy learning? What kind of **asymptotic regret guarantees** can we get?

Simulation Example

Inverse-propensity learning



Efficient policy learning



Here, we took Π to be the set of **depth-2 decision trees**; the optimal treatment boundary is the blue curve. The colors depict average decisions across many simulations

The **policy regret** of IPW was 2.3$\times$ higher than our method's.

Efficient Treatment Effect Estimation

Theorem. (Athey and Wager, 2017) Let $\hat{\pi}_{DR}$ be the maximizer of $\widehat{Q}_{DR}$ over Π, where

$$\widehat{Q}_{DR}(\pi) = \sum_{i=1}^{n} \pi(X_i) \left(\hat{\mu}_+(X_i) - \hat{\mu}_-(X_i) + W_i \frac{Y_i - \hat{\mu}_{W_i}(X_i)}{\hat{e}_{W_i}(X_i)} \right),$$

and let π^* be the best policy in Π. Assume that $\hat{\mu}_\pm(\cdot)$ and $\hat{e}(\cdot)$ are estimated via an $o(n^{-1/4})$-consistent method with cross-fitting.

If Π has a finite VC-dimension, the **regret** of $\hat{\pi}_{DR}$ decays as

$$Q(\pi^*) - Q(\hat{\pi}_{DR}) = \mathcal{O}_P \left(\sqrt{V(\pi^*) \left(1 + \log \left(\frac{V_{\max}}{V(\pi^*)} \right) \right) \frac{VC(\Pi)}{n}} \right),$$

where V_{π^*} is the semiparametric **efficient variance** for estimating $Q(\pi^*)$, and $V_{\max}$ is a bound for $\sup \{ V(\pi) : \pi \in \Pi \}$.

Proof Ingredients: Efficient Coupling

We start by coupling $\widehat{Q}_{DR}$ with the **efficient score estimator**,

$$\widetilde{Q}_{DR}(\pi) = \sum_{i=1}^{n} \pi(X_i) \left(\mu_+(X_i) - \mu_-(X_i) + W_i \frac{Y_i - \mu_{W_i}(X_i)}{e_{W_i}(X_i)} \right).$$

Many classical results in semiparametric analysis (e.g., Newey, 1994) rely on showing that $|\widehat{Q}_{DR}(\pi) - \widetilde{Q}_{DR}(\pi)| = o_P(1/\sqrt{n})$; this usually follows from $o_P(n^{-1/4})$-consistency of $\hat{\mu}_\pm(\cdot)$ and $\hat{e}_\pm(\cdot)$.

Here, we need to go further, and require **uniform coupling**:

$$\sup_{\pi \in \Pi} \left| \widehat{Q}_{DR}(\pi) - \widetilde{Q}_{DR}(\pi) \right| = o_P(1/\sqrt{n}).$$

We use **cross-fitting** (i.e., held-out prediction for $\hat{\mu}_w(X_i)$ and $\hat{e}(X_i)$) as in Chernozhukov et al. (2017) and Schick (1986).

Proof Ingredients: Concentration

Given our coupling result, we can now study **concentration** of

$$\widetilde{Q}_{DR}(\pi) = \sum_{i=1}^{n} \pi(X_i) \left(\mu_+(X_i) - \mu_-(X_i) + W_i \frac{Y_i - \mu_{W_i}(X_i)}{e_{W_i}(X_i)} \right)$$

over the class $\pi \in \Pi$. Defining

$$V(\pi) \leq V_{\max} := \frac{1}{4} \mathbb{E} \left[\left(\mu_+(X_i) - \mu_-(X_i) + W_i \frac{Y_i - \mu_{W_i}(X_i)}{e_{W_i}(X_i)} \right)^2 \right],$$

we can remix Dudley's classical **chaining argument** to verify that

$$\sup \left\{ \left| \widetilde{Q}_{DR}(\pi) - Q_{DR}(\pi) \right| : \pi \in \Pi \right\} = \mathcal{O}_P \left(\sqrt{\frac{V_{\max} \, VC(\Pi)}{n}} \right).$$

The key idea is to bound Rademacher complexity via chaining with a **random distance function** that depends on the data. Going from $V_{\max}$ to $V(\pi^*) \log(V_{\max}/V(\pi^*))$ involves partial chaining.

Summary

We found that **policy regret** is controlled as

$$Q(\pi^*) - Q(\hat{\pi}_{DR}) = \mathcal{O}_P\left(\sqrt{V(\pi^*)\left(1 + \log\left(\frac{V_{\max}}{V(\pi^*)}\right)\right)\frac{VC(\Pi)}{n}}\right).$$

If we just have a single policy π, the **optimal confidence intervals** for the improvement of $\pi(\cdot)$ over the opposite policy $-\pi(\cdot)$ scale as

$$\text{length of conf. interval for } Q(\pi) = \mathcal{O}_P\left(\sqrt{V(\pi)\,/\,n}\right).$$

Ignoring constants and log-factors, our regret bounds scale as $\sqrt{VC(\Pi)}$ times the optimal confidence interval length for π^*.

In the paper, we have **matching lower bounds** for minimax regret, and allow for the **policy dimension** $\mathrm{VC}(\Pi)$ to grow as $o(\sqrt{n})$.

Closing Thoughts

We want a **low regret** (or high Q-value) policy $\hat{\pi} \in \Pi$, with

$$\pi^* := \operatorname{argmax} \{Q(\pi) : \pi \in \Pi\},$$
$$Q(\pi) := 2\mathbb{E}\left[Y(\pi(X)) - \frac{Y(-1) + Y(+1)}{2}\right],$$

The nature of this problem depends on the **structure** of Π.

- If Π is unrestricted, policy learning is a **CATE problem**, and amounts to learning $\tau(x) = \mathbb{E}\left[Y(+1) - Y(-1) \,\middle|\, X = x\right]$.
- If Π is a doubleton, policy learning is a **ATE problem** equivalent to learning $\tau = \mathbb{E}\left[Y(+1) - Y(-1)\right]$.

When Π is structured, we can obtain optimal bounds by considering policy learning as a **continuum of ATE problems**.

Closing Thoughts

When implementing our method in practice, recall that **efficient ATE** estimation involves **consistent CATE** estimation:

$$\widehat{Q}_{DR}(\pi) = \sum_{i=1}^{n} \pi(X_i) \left(\underbrace{\hat{\mu}_+(X_i) - \hat{\mu}_-(X_i)}_{\hat{\tau}(X_i)} + W_i \frac{Y_i - \hat{\mu}_{W_i}(X_i)}{\hat{e}_{W_i}(X_i)} \right).$$

If we use a machine method to estimate the nuisance components, we should consider using one that **emphasizes accuracy of** $\hat{\tau}(\mathbf{x})$, e.g., a **causal forest** or a **jointly estimated lasso**.

Thanks!