Efficient Policy Learning

Machine Learning and Causal Inference, 2017

Based on Athey and Wager (2017)

Drug Facts

Active ingredient (in each tablet) — Purpose

Ranitidine 150 mg (as ranitidine hydrochloride 168 mg).. Acid reducer

Uses

- relieves heartburn associated with acid indigestion and sour stomach
- prevents heartburn associated with acid indigestion and sour stomach brought on by eating and drinking certain foods or beverages

Warnings

Allergy alert: Do not use if you are allergic to ranitidine or other acid reducers

Do not use

- if you have trouble or pain swallowing food, vomiting with blood, or bloody or black stools. These may be signs of a serious condition. See your doctor.
- with other acid reducers
- if you have kidney disease, except under the advice and supervision of a doctor

Ask a doctor before use if you have

- had heartburn over 3 months. This may be a sign of a more serious condition.
- heartburn with **lightheadedness, sweating or dizziness**
- chest pain or shoulder pain with shortness of breath; sweating; pain spreading to arms, neck or shoulders; or lightheadedness
- frequent **chest pain**
- frequent wheezing, particularly with heartburn
- unexplained weight loss
- nausea or vomiting
- stomach pain

Stop use and ask a doctor if

- your heartburn continues or worsens
- you need to take this product for more than 14 days

If pregnant or breast-feeding, ask a health professional before use.

Keep out of reach of children. In case of overdose, get medical help or contact a Poison Control Center right away.

Directions

- adults and children 12 years and over:
  - to **relieve** symptoms, swallow 1 tablet with a glass of water
  - to **prevent** symptoms, swallow 1 tablet with a glass of water **30 to 60 minutes before** eating food or drinking beverages that cause heartburn
  - can be used up to twice daily (do not take more than 2 tablets in 24 hours)
- children under 12 years: ask a doctor

Other information

- do not use if printed foil under bottle cap is open or torn
- avoid excessive heat or humidity
- store at 20º-25ºC (68º-77ºF)
- this product is sodium and sugar free

Inactive ingredients

hypromellose, magnesium stearate, microcrystalline cellulose, synthetic red iron oxide, titanium dioxide, triacetin

Questions? call **1-888-285-9159** (English/Spanish) M – F, 8:30 – 5 EST, or visit **www.zantacotc.com**

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.
- $\ldots$

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 $i = 1, \ldots, n$, 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?

The **optimal policy** is $\pi^* := \text{argmax}\{\mathbb{E}[Y(\pi(X))] : \pi \in \Pi\}$, or,

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

In some cases, policy learning reduces to classical statistical tasks:

- If Π has **no structure**, e.g., $\mathcal{X}$ is discrete, Π contains all $2^{|\mathcal{X}|}$ assignments, finding π^* is just **non-parametric regression**.
- If Π is a **doubleton** $\{\pi_+(x) = +1, \pi_-(x) = -1\}$, then $Q(\pi_+)$ is (half) the average treatment effect; finding π^* reduces to **ATE estimation** in observational studies (Hahn, 1998; Heckman et al., 1998; Hirano et al., 2003; Robins et al., 1995; Rosenbaum, 2002; Rubin, 1974; van der Laan and Rose, 2011;...).
- If Π has **structure**, e.g., if Π consists of linear rules, then... **?**

Policy Learning: Take 1

The **Bayes-optimal** policy is clearly

$$\pi_{\text{bayes}}(x) = 1\left(\{\tau(x) > 0\}\right).$$

Suggests a simple **plug-in** strategy for policy learning:

1. Estimate the CATE function as $\hat{\tau}(\cdot)$, and then
2. Deploy a policy $\hat{\pi}(x) = 1(\{\hat{\tau}(x) > 0\})$.

Given **unconfoundedness** (Rosenbaum & Rubin, 1983),

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i,$$

there are **numerous methods** available for this (Athey & Imbens, 2016; Hill, 2011; Imai & Ratkovic, 2013; Künzel & al., 2017; Powers & al., 2017; Shalit & al., 2017; W. & Athey, 2017).

This plug-in is strategy is simple (and popular)... but does it solve our problem?

Policy Learning: Take 1

Remarkably, in properly specified **non-parametric setups**, this is optimal, and the best strategies for policy learning take the form

$$\hat{\pi}^*(x) = \mathbf{1}\left(\{\hat{\tau}(x) > 0\}\right),$$

where $\hat{\tau}(\cdot)$ is an efficient estimate of $\tau(\cdot)$. In particular,

- **Manski (2004)** considers the case where x is discrete, and studies **conditional empirical success** rules from an asymptotic perspective.
- **Hirano and Porter (2009)** show that, under LeCam **local asymptotics** where effect sizes shrink as $1/\sqrt{n}$, such thresholding rules are optimal in a broad class of problems.
- **Stoye (2009)** derives **exact minimax** rules when x is discrete and $Y_i \,\big|\, X_i = x$ is bounded, and shows that thresholding rules are optimal with matching; with randomization, intriguing small-sample phenomena appear.

Imposing **structure** on Π is essential in many applications (see also Kitagawa & Tetenov, 2015). We use many features with a non-parametric specification to make **unconfoundedness plausible**,

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i.$$

Conversely, the policy $\pi(\cdot)$ must be **implementable in practice**. Features that should not be used in $\pi(\cdot)$ 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** (e.g., Bhattacharya and Dupas, 2012).
- **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**,

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

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

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

This insight, along with the induced policy learner,

$$\hat{\pi}_{IPW} = \operatorname{argmax}\left\{\frac{1}{2}\sum_{i=1}^{n}\frac{\pi(X_i)W_i Y_i}{\mathbb{P}\left[W = W_i \,\middle|\, X = X_i\right]} : \pi \in \Pi\right\},$$

has been independently studied across several fields, including **statistics** (Zhao, Zeng, Rush and Kosorok, 2014), **machine learning** (Beygelzimer and Langford, 2009; Swaminathan and Joachims, 2015), and **economics** (Kitagawa and Tetenov, 2015).

Inverse-Propensity Policy Learning: Pros

The inverse-propensity weighted method uses

$$\hat{\pi}_{IPW} = \operatorname{argmax}\left\{\frac{1}{2}\sum_{i=1}^{n}\frac{\pi(X_i)W_iY_i}{\mathbb{P}\left[W = W_i \,\middle|\, X = X_i\right]} : \pi \in \Pi\right\}.$$

In general, the resulting procedure is **consistent**. Moreover:

- We can establish **policy regret bounds** (Kitagawa and Tetenov, 2015; Swaminathan and Joachims, 2015):

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

- Can be implemented as a **weighted classification problem**:

$$\hat{\pi}_{IPW} = \operatorname{argmax}\left\{\sum_{i=1}^{n}\pi(X_i)\,\operatorname{sign}(\Gamma_i)\,|\Gamma_i| : \pi \in \Pi,\ \Gamma_i := \cdots\right\}.$$

Inverse-Propensity Policy Learning: Cons

The inverse-propensity weighted method uses

$$\hat{\pi}_{IPW} = \operatorname{argmax}\left\{\frac{1}{2}\sum_{i=1}^{n}\frac{\pi(X_i)W_iY_i}{\mathbb{P}\left[W = W_i \,\middle|\, X = X_i\right]} : \pi \in \Pi\right\}.$$

In general, the resulting procedure is **consistent**. However:

- The resulting estimator is not **translation invariant** in Y_i.
- The corresponding **regret bounds** are not translation invariant either.
- . . .

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.

- The goal of this talk is to develop an **efficiency theory** for policy learning.

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.

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 $\pi(\cdot)$ need to 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).

Average treatment effect estimation is **policy learning with a doubleton** Π; now, we let Π be a finite-dimensional continuum.

Efficient Treatment Effect Estimation

Recall that inverse-propensity weighted policy learning uses

$$\hat{\pi}_{IPW} = \operatorname{argmax}\left\{\widehat{Q}_{IPW}(\pi) : \pi \in \Pi\right\}, \; \widehat{Q}_{IPW}(\pi) := \frac{1}{2}\sum_{i=1}^{n} \frac{\pi(X_i) W_i Y_i}{\hat{e}_w(X_i)}.$$

Note that $\widehat{Q}(\pi)$ estimates (half) an average treatment effect,

$$Q(\pi) = \frac{1}{2}\left(\mathbb{E}\left[Y(\pi(X))\right] - \mathbb{E}\left[Y(-\pi(X))\right]\right),$$

where "treated" people get policy $\pi(\cdot)$ and controls get $-\pi(\cdot)$. The efficient estimator for $Q(\pi)$ in this setup is well known in the **semiparametric efficiency** literature (Bickel et al., 1998; Hahn, 1998; Hirano et al., 2003; Robins and Rotnitzky, 1995):

$$\widehat{Q}_{DR}(\pi) = \frac{1}{2}\sum_{i=1}^{n} \pi(X_i)\left(\hat{\mu}_{+}(X_i) - \hat{\mu}_{-}(X_i) + W_i \frac{Y_i - \hat{\mu}_{W_i}(X_i)}{\hat{e}_{W_i}(X_i)}\right).$$

Efficient Treatment Effect Estimation

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

$$\widehat{Q}_{DR}(\pi) = \frac{1}{2}\sum_{i=1}^{n} \pi(X_i)\left(\hat{\mu}_+(X_i) - \hat{\mu}_-(X_i) + W_i \frac{Y_i - \hat{\mu}_{W_i}(X_i)}{\hat{e}_{W_i}(X_i)}\right).$$

We can immediately note the following:

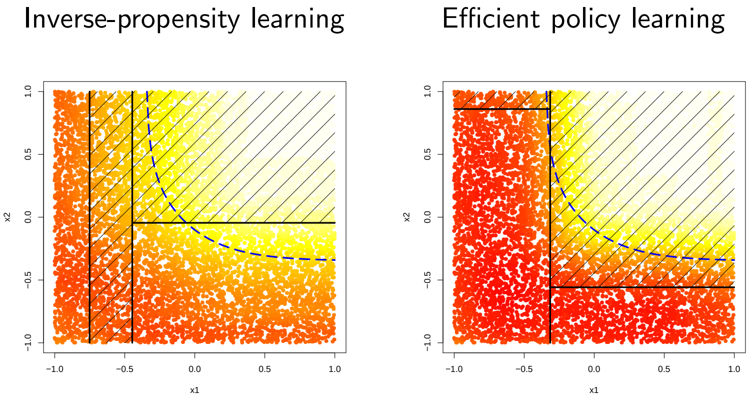
- For a single, deterministic policy π, we know that

$$\sqrt{n}\left(\widehat{Q}_{DR}(\pi) - Q(\pi)\right) \Rightarrow \mathcal{N}\left(0,\ V(\pi)\right),$$
$$4V(\pi) := \operatorname{Var}\left[\pi(X)\left(\mu_+(X) - \mu_-(X)\right)\right]$$
$$+\,\mathbb{E}\left[\operatorname{Var}\left[Y(-)\,\middle|\,X\right] / e_-(x)\right] + \mathbb{E}\left[\operatorname{Var}\left[Y(+)\,\middle|\,X\right] / e_+(x)\right],$$

  and this is the **efficient asymptotic variance**.
- Again, can be implemented via **weighted classification**.

What kind of **asymptotic regret guarantees** can we get?

Simulation Example



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

$$\widehat{Q}_{DR}(\pi) = \frac{1}{2}\sum_{i=1}^{n} \pi(X_i)\left(\hat{\mu}_{+}(X_i) - \hat{\mu}_{-}(X_i) + W_i \frac{Y_i - \hat{\mu}_{W_i}(X_i)}{\hat{e}_{W_i}(X_i)}\right),$$

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.

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

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

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

Discussion

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

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

If we just has 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.

- Very **heuristically**, if we think of optimizing over a class of dimensions $VC(\Pi)$ as picking the best of roughly $2^{VC(\Pi)}$ policies, this is essentially the best result we could hope for.