A Survey on Policy Learning

MATH 818.01 Midterm Survey

Wonjun Choi

October 21, 2020

Abstract

This paper surveys growing literature on policy learning in the interdisciplinary area of economics, statistics, and computer science. Policy learning incorporates statistical decision making, AI/ML algorithms, and potential outcome model in economics literature to find optimal policy assignment rule. Policy learning predicts the expected outcome of policy given practical restrictions such as budget constraint, fairness, or political reasons. Considerations raised in the economics literature, applications of AI/ML algorithms, and mathematical foundation of the results are briefly introduced in turn. A suggestion for the final project is also presented in the final section.

1 Introduction

Consider a multi-armed bandit problem that we have learned in our class. Suppose there are five restaurants $D = 1, ..., 5$ each of which sells a sandwich that gives random utility $Y(D)$, respetively. Suppose you are a dean of the department and you plan to address a lunch meeting with a hundred undergraduate students $i = 1, 2, ..., 100$. Now you are interested in how to maximize the expected total welfare for of the lunch meeting.

One strategy you can deploy is to allocate ten graduate students $j = 1, ..., 10$ with feature X_j to each restaurants and observe their utility Y_j's. By making the utility function $Y_j(D)$ of each sandwiches as a function of X_j, you can derive the expected total welfare $\sum_{i=1}^{100} Y_i(D_i)$ of your policy $p : \mathcal{X} \to \{1, 2, ..., 5\}$ under some very restrictive assumptions such as X_i and X_j have the same distribution.

This example resembles the problem in which economists, as well as numerous researchers in various disciplines, are interested. They want to find a guide for designing poilcies from the data of the past experiences. Approaches that researchers prefer to may vary on their disciplines. In this paper, I restrict the focus to the literature on the policies, most likely to be implemented by the government, that policymakers might be interested in.

Specifically, this paper assumes the *offline* setup for the data collection scheme. Unlike the online setup where data can be collected through online experiments, policimakers usually only have access

to 'observational' data. Moreover, collecting data through experiment is oftenly hindered by ethical or political restrictions.

I also restrict the focus of this paper in Frequentist's point of view. Different perspectives of Frequentists and Bayesians have a long history in statistical literature. Since Friquentist's perspective has a relative advantage for an inference purpose which is also a great concern of policimakers, I stick on this perspective in this survey.

The rest of the paper consists as follows. Section 2 introduces frameworks that economists use for policy evaluation. Section 3 is about the policy learning literature that adopts AI/ML methods for estimating and optimizing policy decision. Section 4 gives a mathematical foundation of the results found in Section 2 and 3. A suggestion for the final project is presented in the final section.

2 Economic Modeling of the Allocation Problem

It is worth to stop and introduce typical approaches in economics for policy evaluation. Introduced notations would ease our conversation throughout this paper.

Policy evaluation has been one of the most important topics in the field of economics. As the government implements various economic policies, the outcomes of the interventions need to be analyzed and have been studied in the framework of *potential outcome*, which was frequently referred to as *Rubin's causal model*.

Potential Outcome Framework

Consider we are giving aspirins to patients and observing their body temperatures. After treatments are assigned to each patient, we can only observe only *one side* of the outcome; the temperature with or without taking aspirin. If one takes her aspirin, we cannot know what the temperature would have been without taking it, and vice versa. This is the fundamental problem of *treatment effect* analysis.

What would be the effect of aspirin on body temperature? It would be the difference of body temperature between taking aspirin and not. With D equals 1 if the treatment is applied and 0 otherwise, let the potential outcome be $Y(D)$, then $Y(1)$ be the outcome with the treatment and $Y(0)$ be the outcome without the treatment. Now we can denote the treatment effect τ as

$$\tau = Y(1) - Y(0).$$

τ cannot be obtained without further assumptions since one of the outcomes is not observed. In economics/statistics literature, the unobserved outcome is called *counterfactual*. I refer [Imbens and Rubin, 2015], among many others, for detailed explanations and issues in treatment effect analysis.

Regret Function and Optimal Policy

One practical concern in designing policy could be 'How should we assign (limited) treatments for the best outcome?'. [Manski, 2004] introduced a framework for this analysis to economists based on

statistical decision rule of statistics literature([Wald, 1950]).

The concern of statistical treatement rule is that how to assign a treatment D to the population based on their covariates(features) $X \in \mathcal{X}$ to maximize utilitarian welfare U[1]. Let's call this assignment rule(function) as *policy* $\pi : \mathcal{X} \to \{0, 1\}$ and the collection of possible policies as $\Pi = \{\pi : \text{some restirctions on } \pi\}$.

It is not difficult to find this kind of problem in reality. For example, consider Youtube's recommendation algorithm. For a given video A, the algorithm decides whether to recommend this video to you or not based on your characteristics[2]. Now let's say Youtube has decided to recommend this video to total 100 people among its users. Then Youtube would want to find an algorithm(policy) to maximize its total viewership(utilitarian welfare) among possible policies.

How can we measure the successfulness of a given policy? In what sense, the optimal policy can be regarded as the best? As already mentioned, we are maximizing the utilitarian welfare so the best policy could be thought as

$$\pi_{opt} = \arg\max_{\pi \in \Pi} U(\pi)$$

where $U(\pi)$ is the total welfare of the population when the policy π is implemented. By defining a *regert function* $R(\pi)$ as a difference between the welfare with the policy π and the best possible outcome,

$$R(\pi) = U(\pi_{opt}) - U(\pi),$$

we can measure the successfulness of the policy π by comparing $R(\pi)$. Now our objective becomes clear: finding a policy funtion π that minimizes the functional $R(\pi)$.

Theoretical efforts have been made to bound a reget function. With a properly decided policy, we can find a uniform bound of the regret function. Notice that if a regret function degenerates fastly enough, we might be willing to adopt that that policy as a solution to our decision making problem. For some results regarding a regret bound and the minimax decision criteria, refer [Manski, 2004] [Stoye, 2009], and [Hirano and Porter, 2009].

3 Policy Learning Embedding AI/ML Algorithms

A similar context can be found in computer science literature; *multi-armed bandit* and their close cousins. In this section, I focus on the offline setup of the problem as economists usually have 'observation' data.

[Dudik et al., 2011] proposes a *doubly robust* estimator for policy evaluation and optimization(learning). As mentioned in the previous section, the treatment effect of a policy cannot be estimated without further assumptions for the missing data problem(counterfactual). There are two typical approaches that the authors refer to as *direct method*(DM) and *inverse propensity score*(IPS), respectively. For a detailed treatment of these methods, refer to their paper or [Imbens and Rubin, 2015]. I briefly

[1] If U is stochastic, consider the mean $E(U)$.

[2] Of course, if we set a policy as $\pi : X \to \{A, B, C, D, ...\}$, we can consider a more complexed decision making.

introduce their experimental setups which are more relevant to our remaining discussions. Notations and explanations are mostly theirs([Dudik et al., 2011]).

Consider i.i.d. data drawn from a distribution D: $(x, c) \sim D$, where $x \in \mathcal{X}$ is the feature vector and $c \in C = \{1, 2, .., , k\}$ is the class label. An action a is chosen by the policy $p(a|x, h)$, where h is the history of previous observations. A reward r_a is revealed while other potential rewards $r_{a'}$ remain unknown. Defining the *value* of a policy π as

$$V^{\pi} = E_{x,\vec{r}}[r_{\pi(x)}],$$

our policy learing objective is to find a policy that maximizes the value function.

[Dudik et al., 2011] provides simulation results by transforming a classical classification problem[3] into a policy learning framework. They consider (x, c) as a observed sample and let (potential) loss as $(x, l_1, l_2, \cdots, l_k)$ with $l_a = 1[a \neq c]$ where $1[\cdot]$ is an indicator function. Then the two problems become identical. The opposite way of transforming is also interesting: we can use an optimization tool for classification to solve our policy optimization.

[Kitagawa and Tetenov, 2018] uses [Dudik et al., 2011]'s IPS estimator and proves that their *empirical welfare maximize* method meets semiparametric efficient minimax regret bound under some assumptions that are commonly used in economics literature. For readers who are interested in the assumptions that economists use may refer to their paper.

As the flexible and powerful features of various AI/ML algorithms remain attractive to economists' eyes, there have been trials to embrace those methods for policy evaluation. Among many others, I introduce [Athey and Wager, 2017] whose method can be equipped with AI/ML techniques. They suggest an algorithm to find such optimal policies:

1. Estimate the potential oucome equation $\hat{m}$ and the some function $\hat{g}$ using any methods whose rate of convergence is known.

2. Construct a score function for the value function $\hat{\Gamma}$ using nuisance components estimated in 1.

3. Find $\hat{\pi} = \arg\max\{\sum(2\pi(X_i) - 1)\hat{\Gamma}\}$ where $\pi(\cdot)$ is a trained weighted classifier.

For detailed instruction for $\hat{m}$, $\hat{g}$, $\hat{\Gamma}$, refer their paper. Here, part 1 and 3 of the algorithm can be obtained with AI/ML methods.

For the part 1, we can use any methods whose rate of convergence in mean square error(MSE) is known. As many statistical(machine) learning techniques use a MSE criteria for their loss functions, various ML estimators can be used for the part 1. Also their convergence rate has been widely investigated nowadays. While optimization in the part 3 is not a convex optimization, the computation of the part 3 could be troublesome. As mentioned before, the problem can be translated into a weighted classification problem so we can use techniques developed for those classifications([Athey and Wager, 2017]).

[3] In a classification problem, we are searching for a classifier $\pi : \mathcal{X} \to C$ that minimizes the classification error.

4 Mathematical/Statistical Foundations

Theoretical properties of aforementioned methods rely on the theory of *uniform convergence* of stochastic(empirical) processes. Deviation of an estimator from the parameter can be bounded with a calculation of *Rademacher Complexity* which is in turn bounded by *Vapnik-Chervonenkis dimesion*(VC-dimension). I largely refers to [Wainwright, 2019]'s descriptions in this section.

Theorem 1 (Glivenko-Cantelli;[Wainwright, 2019] p.100)**.** *For any distribution, the empirical CDF $\hat{F}_n$ is a strongly consistent estimator of the population CDF in the uiform norm, meaning that*

$$||\hat{F}_n - F||_\infty \to 0 \quad a.s.$$

where $||\cdot||_\infty$ is the supremum norm.

Statistical application of this theorm often involves considering an estimator as a function of (empirical) distribution. As an empirical distribution converges to true distribution, we can prove some results by giving some restrictions on the class of functionals(estimators).

Rademacher complexity is a kind of measurement of the complexity of a function set. In our problem, we are searching for the optimal policy π within the collection of feasible policies Π. Thus, our result depends on the complexity of the collenction Π. Consider a set of functions $\mathcal{F}$. Fix a point $x_1^n = (x_1, x_2, ..., x_n)$ and define $\mathcal{F}(x_1^n)$ as

$$\mathcal{F}(x_1^n) = \{f(x_1), f(x_2), ..., f(x_n) : f \in \mathcal{F}\}$$

with a *Rademacher variable* ϵ which takes the value 1 and -1 equiprobably, the *empirical Rademacher complexity* is given by

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) = E_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right| \right].$$

Now, considering x_1^n as a realization of a random vector X_1^n, we gain the definition of Rademancher complexity of the function set $\mathcal{F}$.

Definition 2 (Rademacher complexity; [Wainwright, 2019] p.104)**.** *With X_1^n, $\mathcal{F}(X_1^n)$, ϵ_i defined as above, the Rademancher complexity of a set $\mathcal{F}$ is the deterministic quantity*

$$\mathcal{R}_n(\mathcal{F}) = E_{X,\epsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| \right].$$

The theorem below shows a connection between aforementioned two ideas for a special case when $\mathcal{F}$ is a set of (uniformly) bounded functions.

Theorem 3 ([Wainwright, 2019] p.105)**.** *For any b-uniformly bounded class of functions $\mathcal{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have*

$$||P_n - P||_\mathcal{F} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta$$

with P-probability at least $1 - \exp(-\frac{n\delta^2}{2b^2})$.
Consequently, as long as $\mathcal{R}_n(\mathcal{F}) = o(1)$, we have $||P_n - P||_\mathcal{F} \to 0 \quad a.s.$

Here, $||P_n - P||_{\mathcal{F}}$ is defined as $\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E_P(f(X))\right|$.

As can be seen in the Theorem 3, we can obtain uniform convergence by properly control Rademancher complexity. Among many methods, one popular approach is to introduce Vapnik-Chervonenkis dimension.

Definition 4 (Shattering and VC-dimension; [Wainwright, 2019] p.112.)**.** *Given a class $\mathcal{F}$ of binary-valued functions, we say that the set $x_1^n = (x_1, ..., x_n)$ is shattered by $\mathcal{F}$ if $card(\mathcal{F}(x_1^n)) = 2^n$. The VC dimension $v(\mathcal{F})$ is the largest integer n for which there is some collection x_1^n of n points that is shattered by $\mathcal{F}$.*

As it is too lengthy a discussion for this survey to introduce a calculation of VC dimension and the related corollaries, I hereafter refer [Wainwright, 2019] for further theories. For classical approaches in controlling the complexity of a set I refer [Van Der Vaart and Wellner, 1996, Van der Vaart, 2000] as well. Many of the properties of AI/ML methods also rely on Rademancher complexity and VC dimension of their set of loss functions. For this, refer [Mohri et al., 2018].

5 Discussion/Conclusion

In the survey, I introduced a problem of treatment assignment rule and one path of the recent development.

While the literature is fastly growing, there are still many unsolved questions remaining. First, in the real application of the problem, using the introduced methods requires several selections from the user. For example, to estimate the part 1 of [Athey and Wager, 2017]'s algorithm, one have to decide(or ensemble) which algorithm to use. Moreover, if the rate of convergence of certain method is not known, one might have to derive it. The part 3 of their algorithms is also computationally tricky. Depending on the numerical method, the result might vary.

Developed models so far are still parsimonious for many applications. For example, as in MAB, a treatment assignment could be repeated during the process. Also, the data collection can be augmented during the process(online setup). Using a longitudinal data might need more modifications especially when using ML methods. Bayesian approach also has a rich literature on the prediction side of this problem.

For the final project, I would like to study the case of 'Disaster support' for COVID-19 in Korea. There was an debate on the way to distribute this aid: whether to provide it for everyone or low-incomed. I would like to investigate on this debate in terms of the consumption stimuli effect as many countries around the world considered similar policy as a fiscal stimuli for economy. Specifically I would like to adopt [Kitagawa and Tetenov, 2018]'s method or [Athey and Wager, 2017]'s method which involves an optimization using AI/ML algorithms.

References

[Athey and Wager, 2017] Athey, S. and Wager, S. (2017). Policy Learning with Observational Data. *arXiv*.

[Dudik et al., 2011] Dudik, Langford, and Li (2011). Doubly Robust Policy Evaluation and Learning.

[Hirano and Porter, 2009] Hirano, K. and Porter, J. R. (2009). Asymptotics for Statistical Treatment Rules. *Econometrica*, 77(5):1683–1701.

[Imbens and Rubin, 2015] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[Kitagawa and Tetenov, 2018] Kitagawa, T. and Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2):591–616.

[Manski, 2004] Manski, C. F. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246.

[Mohri et al., 2018] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

[Stoye, 2009] Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81.

[Van der Vaart, 2000] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[Van Der Vaart and Wellner, 1996] Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

[Wainwright, 2019] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

[Wald, 1950] Wald, A. (1950). Statistical decision functions.