Econometrica, Vol. 89, No. 1 (January, 2021), 133–161

POLICY LEARNING WITH OBSERVATIONAL DATA

SUSAN ATHEY
Stanford Graduate School of Business, Stanford University

STEFAN WAGER
Stanford Graduate School of Business, Stanford University

In many areas, practitioners seek to use observational data to learn a treatment assignment policy that satisfies application-specific constraints, such as budget, fairness, simplicity, or other functional form constraints. For example, policies may be restricted to take the form of decision trees based on a limited set of easily observable individual characteristics. We propose a new approach to this problem motivated by the theory of semiparametrically efficient estimation. Our method can be used to optimize either binary treatments or infinitesimal nudges to continuous treatments, and can leverage observational data where causal effects are identified using a variety of strategies, including selection on observables and instrumental variables. Given a doubly robust estimator of the causal effect of assigning everyone to treatment, we develop an algorithm for choosing whom to treat, and establish strong guarantees for the asymptotic utilitarian regret of the resulting policy.

KEYWORDS: Double robustness, empirical welfare maximization, minimax regret, semiparametric efficiency.

1. INTRODUCTION

THE PROBLEM OF LEARNING treatment assignment policies or mappings from individual characteristics to treatment assignments is ubiquitous in applied economics and statistics. It arises, for example, in medicine when a doctor must decide which patients to refer for a risky surgery; in marketing when a company needs to choose which customers to send targeted offers to; and in government and policy settings, when assigning students to educational programs or inspectors to buildings and restaurants.

The treatment assignment problem rarely arises in an unconstrained environment. Treatments are often expensive, and so a policy may need to respect budget constraints. Policies may need to be implemented in environments characterized by human or machine constraints; for example, emergency medical professionals or police officers may need to implement decision policies in the field, where a simple decision tree might be used. For internet or mobile services, algorithms may need to determine the set of information displayed to a user very quickly, and a simple lookup table may decrease the time it takes to respond to a user's request. Fairness constraints may require a treatment assignment policy to depend only on particular

Susan Athey: athey@stanford.edu
Stefan Wager: swager@stanford.edu

We circulated an earlier draft of this paper under the title "Efficient Policy Learning"; the current title was chosen following guidance from the review process. We are grateful for helpful conversations with colleagues including Victor Chernozhukov, David Hirshberg, Guido Imbens, Michael Kosorok, Alexander Luedtke, Eric Mbakop, Whitney Newey, Xinkun Nie, Molly Offer-Westort, Alexander Rakhlin, James Robins, Erik Sverdrup, Max Tabord-Meehan and Zhengyuan Zhou, and for feedback from the editor, referees, as well as seminar participants at a variety of universities and workshops. We thank Guido Imbens for sharing the GAIN dataset with us. Financial support was provided by the Sloan Foundation, Office of Naval Research grant N00014-17-1-2131, National Science Foundation grant DMS-1916163, and a Facebook Faculty Award.

 https://doi.org/10.3982/ECTA15732

types of covariates (e.g., test scores or income), even when other covariates are observed.

This paper is about using observational data to learn policies that respect the types of constraints outlined above. The existing literature on policy learning has mostly focused on the setting where we want to optimize allocation of a binary treatment using data from a randomized trial, or from a study with a known, random treatment assignment policy. In many problems, however, one may need to leverage richer forms of observational data to learn treatment assignment rules. For example, if we want to learn whom to prescribe a drug to based on data from a clinical trial, we need to have methods that deal with noncompliance and resulting endogenous treatment assignments.[1] Or, if we are interested in offering some customers discounts, then we need methods that let us study interventions to continuous variables (e.g., price) rather than just discrete ones. The goal of this paper is to develop methods for policy learning that do not just work in randomized trials (or related settings), but can instead work with a rich variety of observational designs.

Formally, we study the problem where we have access to observational data and want to use it to learn a policy that maps a subject's characteristics $X_i \in \mathcal{X}$ to a binary decision, $\pi : \mathcal{X} \to \{0, 1\}$. The practitioner has also specified a class Π that encodes problem-specific constraints pertaining to budget, functional form, fairness, etc., and requires that our learned policy $\hat{\pi}$ satisfies these constraints, $\hat{\pi} \in \Pi$. Then, following Manski (2004, 2009), Hirano and Porter (2009), Stoye (2009, 2012), and Kitagawa and Tetenov (2018), we seek guarantees on the regret $R(\hat{\pi})$, that is, the difference between the expected utility from deploying the learned policy $\hat{\pi}$ over a target population and the best utility that could be achieved from deploying any policy in the class Π over the population.

Our paper builds on a rich literature at the intersection of econometrics, statistics, and computer science on learning structured treatment assignment rules, including Kitagawa and Tetenov (2018), Swaminathan and Joachims (2015), and Zhao, Zeng, Rush, and Kosorok (2012). Most closely related to us, Kitagawa and Tetenov (2018) studied a special case of our problem where treatments are binary and exogenous with known assignment probabilities, and showed that an algorithm based on inverse-probability weighting achieves regret that depends optimally on the sample size and the complexity of the policy class Π.[2]

Here, we develop a new family of algorithms that achieve regret guarantees with optimal dependence on sample size and on Π, but under considerably more generality on the sampling design. We consider both the classical case where we want to optimize a binary treatment, and a related setting where we want to optimize infinitesimal nudges to a continuous treatment (e.g., a price). Moreover, our approach can leverage observational data where the treatment assignment mechanism may either be exogenous with unknown assignment probabilities, or endogenous, in which case we require an instrument.

Our approach starts from recent unifying results of Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2016) on semiparametrically efficient estimation. As discussed in more detail in Section 2.1, Chernozhukov et al. (2016) showed that in many

[1] If we believed that compliance patterns when we deploy our policy would be similar to those in the clinical trial, then an intent-to-treat analysis may be a reasonable way to side-step endogeneity concerns. However, if we suspect that compliance patterns may change (e.g., if patients may be more likely to adhere to a treatment regime prescribed by their doctor than one randomly assigned in a clinical trial), then using an analysis that disambiguates received treatment from assigned treatment is necessary.

[2] Kitagawa and Tetenov (2018) also considered the case where treatment assignment probabilities are unknown; in this case, however, their method no longer achieves optimal dependence on the sample size.

problems of interest, we can construct efficient estimates of average-treatment-effect-like parameters θ as

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_i, \tag{1}$$

where $\widehat{\Gamma}_i$ is an appropriate doubly robust score for the target estimand under the intervention of interest. This approach can be used to target the average effect of a binary treatment, the average derivative of a continuous treatment, and other related estimands.

In this paper, we find that whenever one can estimate the average utility of treating everyone[3] using an estimator of the type (1) built via the doubly robust construction of Chernozhukov et al. (2016), we can also usefully learn whom to target with the intervention via a simple procedure: Given a prespecified policy class Π (e.g., linear decision rules or finite-depth decision trees), we propose using the treatment assignment rule $\hat{\pi}$ that solves[4]

$$\hat{\pi} = \operatorname{argmax}\left\{ \frac{1}{n}\sum_{i=1}^{n} \big(2\pi(X_i) - 1\big)\widehat{\Gamma}_i : \pi \in \Pi \right\}, \tag{2}$$

where $\widehat{\Gamma}_i$ are the same doubly robust scores as used in (1). Our main result is that, under regularity conditions, the resulting policies $\hat{\pi}$ have regret $R(\hat{\pi})$ bounded on the order of $\sqrt{\mathrm{VC}(\Pi)/n}$ with high probability. Here, $\mathrm{VC}(\Pi)$ is the Vapnik–Chervonenkis dimension of the class Π and n is the sample size. We also highlight how the constants in this bound depend on fundamental quantities from the semiparametric efficiency literature.

Our proof combines results from semiparametrics with carefully tailored analysis tools that build on classical ideas from empirical process theory. The reason we obtain strong guarantees for the approach (2) is closely tied to robustness properties of the estimator (1). In the setting where we only want to estimate a single average effect parameter, it is well known that nondoubly robust estimators can also be semiparametrically efficient (Hirano, Imbens, and Ridder (2003)). Here, however, we need convergence results that are strong enough to withstand optimization over the whole class Π. The fact that doubly robust estimators are fit for this task is closely related to their ability to achieve semiparametric efficiency under general conditions, even if nuisance components are estimated via black-box machine learning methods for which we can only guarantee fast enough convergence in mean-squared error (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), van der Laan and Rose (2011)).

We spell out our general framework in Section 2. For intuition, however, it is helpful to first consider this approach in the simpler case where we want to study the effect of a binary treatment $W_i \in \{0, 1\}$ on an outcome $Y_i \in \mathbb{R}$ interpreted as a utility and are willing to assume selection on observables (unconfoundedness): We have potential outcomes $\{Y_i(0), Y_i(1)\}$ such that $Y_i = Y_i(W_i)$ and $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i | X_i$ (Imbens and Rubin (2015)). Then the utilitarian regret of deploying a policy $\pi \in \Pi$ is (Manski (2009))

$$R(\pi) = \max\big\{\mathbb{E}\big[Y_i\big(\pi'(X_i)\big)\big] : \pi' \in \Pi\big\} - \mathbb{E}\big[Y_i\big(\pi(X_i)\big)\big], \tag{3}$$

[3]Throughout this paper, we assume that there is no interference, that is, assigning one unit to treatment does not affect outcomes for others. For a discussion of treatment effect estimation under interference, see Hudgens and Halloran (2008), Manski (2013), and references therein.

[4]If this optimization problem has multiple solutions, we set $\hat{\pi}$ to an arbitrary maximizer of the objective. Our formal results apply simultaneously to all solutions of (2).

and we can construct our estimator (2) using the well-known augmented inverse-propensity weighted scores of Robins, Rotnitzky, and Zhao (1994),[5]

$$\widehat{\Gamma}_i = \hat{m}(X_i, 1) - \hat{m}(X_i, 0) + \frac{W_i - \hat{e}(X_i)}{\hat{e}(X_i)\big(1 - \hat{e}(X_i)\big)}\big(Y_i - \hat{m}(X_i, W_i)\big),$$
$$e(x) = \mathbb{P}[W_i = 1 | X_i = x], \qquad m(x, w) = \mathbb{E}\big[Y_i(w) | X_i = x\big], \tag{4}$$

where $\hat{e}(x)$ and $\hat{m}(x, w)$ denote nonparametric estimates of $e(x)$ and $m(x, w)$, respectively. In this setup, our result implies that—under regularity conditions—the estimator (2) with scores (4) has regret (3) bounded on the order of $\sqrt{\mathrm{VC}(\Pi)/n}$.

Even in this simplest case, our result is considerably stronger than results currently available in the literature. The main result of Kitagawa and Tetenov (2018) is that, if treatment propensities $e(X_i)$ are known, then a variant of inverse-propensity weighted policy learning achieves regret on the order of $\sqrt{\mathrm{VC}(\Pi)/n}$. However, in observational studies where the treatment propensities are unknown, the bounds of Kitagawa and Tetenov (2018) depend on the rate at which we can estimate $e(\cdot)$, and will generally decay slower than $1/\sqrt{n}$. The only other available $1/\sqrt{n}$-bounds for policy learning in observational studies with a binary treatment that we are aware of are a result of van der Laan, Dudoit, and van der Vaart (2006) for the case where Π consists of a finite set of policies whose cardinality grows with n, and a result of Kallus (2018) in the special case $m(\cdot, w)$ is assumed to belong to a reproducing kernel Hilbert space. The idea of using doubly robust scores to learn optimal treatment assignment of a binary treatment has been previously discussed in Dudík, Langford, and Li (2011) and Zhang, Tsiatis, Davidian, Zhang, and Laber (2012); however, neither paper provides a regret bound for this approach.

In the more general case where the observed treatment assignments W_i may be continuous and/or we may need to use instrumental variables to identify causal effects, both the methods and regret bounds provided here are new. By connecting the policy learning problem to the semiparametric efficiency literature, we are able to develop a general framework that applies across a variety of settings.

1.1. *Related Work*

The literature on optimal treatment allocation has been rapidly expanding across several fields. In the econometrics literature, the program of learning regret-optimal treatment rules was started by Manski (2004, 2009). One line of work considers the case where the policy class is unrestricted, and the optimal treatment assignment rule simply depends on the sign of the conditional average treatment effect for each individual unit. In this setting, Hirano and Porter (2009) showed that when $1/\sqrt{n}$-rate estimation of the conditional average treatment effect function is possible, then treatment assignment rules obtained by thresholding an efficient estimate of the conditional average treatment effect are asymptotically minimax-optimal. Meanwhile, Stoye (2009) derived finite sample minimax decision rules in a class of problems where both the response surfaces and the policies π may depend arbitrarily on covariates. Further results are given in Armstrong and Shen (2013), Bhattacharya and Dupas (2012), Chamberlain (2011), Dehejia (2005), Kasy (2016), Stoye (2012), and Tetenov (2012).

[5]See Section 5.1 for a detailed discussion of how to implement our policy learner (2) based on these augmented inverse-propensity weighted scores in practice.

Building on this line of work, Kitagawa and Tetenov (2018) studied policy learning in a nonparametric setting where the learned policy $\hat{\pi}$ is constrained to belong to a structured class Π and show that, in this case, we can obtain regret bounds relative to the best policy in Π that scale with the complexity of the class Π. A key insight from Kitagawa and Tetenov (2018) is that, when propensity scores are known and Π has finite VC dimension, it is possible to get $1/\sqrt{n}$-rate regret bounds for policy learning over a class Π even if the conditional average treatment effect function itself cannot be estimated at a $1/\sqrt{n}$-rate; in other words, we can reliably find a nearly best-in-class policy without needing to accurately estimate a model that describes all causal effects. As discussed above, our paper builds on this work by considering rate-optimal regret bounds for best-in-class policy learning in observational studies where propensity scores are unknown and treatment assignment may be endogenous, etc.

One difference between our results and those of Kitagawa and Tetenov (2018) is that the latter provide finite sample regret bounds, whereas our results are asymptotic in the sample size n. The reason for this is that our bounds rely on results from the literature on semiparametric estimation (Bickel, Klaassen, Ritov, and Wellner (1998), Chernozhukov et al. (2016), Chen, Hong, and Tarozzi (2008), Hahn (1998), Newey (1994), Robins and Rotnitzky (1995)), which themselves are asymptotic. Recently, Armstrong and Kolesár (2017) showed that, in a class of average treatment effect estimation problems, finite sample conditionally minimax linear estimators are asymptotically efficient, thus providing a connection between desirable finite sample guarantees and asymptotic optimality. It would be interesting to examine whether similar connections are possible in the policy learning case.

Policy learning from observational data has also been considered in parallel literatures developed in both statistics (Luedtke and van der Laan (2016), Qian and Murphy (2011), Zhang et al. (2012), Zhao et al. (2012)) and machine learning (Beygelzimer and Langford (2009), Dudík, Langford, and Li (2011), Kallus (2018), Swaminathan and Joachims (2015)). Two driving themes behind these literatures are the development of performant algorithms for solving the empirical maximization problems (and relaxations thereof) that underlie policy learning, and the use of doubly robust objectives for improved practical performance. Kallus (2018), Swaminathan and Joachims (2015), and Zhao et al. (2012) also proved regret bounds for their methods; however, they do not achieve a $1/\sqrt{n}$ sample dependence, with the exception of Kallus (2018) in the special case of the reproducing kernel Hilbert space setting described above. Finally, Luedtke and Chambaz (2020) proposed a class of regret bounds that decay faster than $1/\sqrt{n}$ by exploiting nonuniform asymptotics; see Section 4 for a further discussion.

The problem of optimal treatment allocation can also be seen as a special case of the broader problem of optimal data-driven decision making. From this perspective, our result is related to the work of Ban and Rudin (2019) and Bertsimas and Kallus (2020), who study data-driven rules for optimal inventory management and related problems. Much like in our case, they advocate learning with a loss function that is directly tied to a utility-based criterion. Finally, we note a growing literature on estimating conditional average treatment effects, including Athey and Imbens (2016), Athey, Tibshirani, and Wager (2019), Nie and Wager (2020), and references therein. Although the goal is similar to that of learning optimal treatment assignment rules, the specific results themselves differ; they focus on squared-error loss rather than utilitarian regret.

2. FROM EFFICIENT POLICY EVALUATION TO LEARNING

Our goal is to learn a policy $\pi \in \Pi$ that maps a subject's features $X_i \in \mathcal{X}$ to a treatment decision: $\pi : \mathcal{X} \to \{0, 1\}$. In order to do so, we assume that we have independent and identically distributed samples (X_i, Y_i, W_i, Z_i), where $Y_i \in \mathbb{R}$ is the outcome we want to intervene on, W_i is the observed treatment assignment, and Z_i is an (optional) instrument used for identifying causal effects. In cases where W_i is exogenous, we simply take $Z_i = W_i$. Throughout our analysis, we interpret Y_i as the utility resulting from our intervention on the ith sample, for example, Y_i could measure the benefit accrued by a subject minus a potentially personalized cost of treatment (in Section 5.1 we demonstrate inclusion of linear costs in the context of an application). We then seek policies that make the expected value of Y_i large.

We define the causal effect of the intervention $\pi(\cdot)$ in terms of the potential outcomes model (Neyman (1923), Rubin (1974)), whereby the $\{Y_i(w)\}$ correspond to utilities we would have observed for the ith sample had the treatment been set to $W_i = w$, and $Y_i = Y_i(W_i)$. When instruments are present, we always assume that the exclusion restriction holds so that this notation is well specified. We consider both examples with a binary treatment $W_i \in \{0, 1\}$ and with a continuous treatment $W_i \in \mathbb{R}$.

In the case where W_i is binary, we follow the existing literature (Hirano and Porter (2009), Kitagawa and Tetenov (2018), Manski (2004), Stoye (2009)), and study interventions that directly specify the treatment level. In this case, the utility of deploying a policy $\pi(\cdot)$ relative to treating no one is (Manski (2009))

$$V(\pi) = \mathbb{E}\big[Y_i\big(\pi(X_i)\big) - Y_i(0)\big], \tag{5}$$

and the corresponding policy regret relative to the best possible policy in the class Π is

$$R(\pi) = \max\big\{V\big(\pi'\big) : \pi' \in \Pi\big\} - V(\pi). \tag{6}$$

As discussed in the Introduction, in this binary setting, Kitagawa and Tetenov (2018) showed that if W_i is exogenous with known treatment propensities, then we can use inverse-propensity weighting to derive a policy $\hat{\pi}$ whose regret $R(\hat{\pi})$ decays as $1/\sqrt{n}$, with

$$\hat{\pi}_{\text{IPW}} = \operatorname{argmax}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1\big(\big\{W_i = \pi(X_i)\big\}\big)Y_i}{\mathbb{P}\big[W_i = \pi(X_i)|X_i\big]} : \pi \in \Pi\right\}. \tag{7}$$

Here, we develop methods that can also be used in observational studies where treatment propensities may be unknown, and where we may need to use instrumental variables to identify $V(\pi)$ from (5).

Meanwhile, when W_i is continuous, we study infinitesimal interventions on the treatment level motivated by the work of Powell, Stock, and Stoker (1989). We define the utility of such an infinitesimal intervention as

$$V(\pi) = \left[\frac{d}{d\nu}\mathbb{E}\big[Y_i\big(W_i + \nu\pi(X_i)\big)\big]\right]_{\nu=0}, \tag{8}$$

and then define regret in terms of $V(\pi)$ as in (6). One interesting conceptual difference that arises in this case is that now our interventions $\pi(X_i) \in \{0, 1\}$ and observed treatment assignments $W_i \in \mathbb{R}$ may take values in different spaces. This can arise, for example, if we want to target customers with personalized discounts and have access to past prices W_i

that take on a continuum of values, but are restricted to considering a class of interventions that only allow us to make a binary decision $\pi(X_i) \in \{0, 1\}$ on whether to offer each customer a small discount or not. The fact that we can still learn low-regret policies via the simple strategy (2) even when these two spaces are decoupled highlights the richness of the policy learning problem.[6]

With both binary and continuous treatments, the regret of a policy π can be written in terms of a conditional average treatment effect function,

$$\tau(x) = \mathbb{E}\big[Y_i(1) - Y_i(0) | X_i = x\big] \text{ or } \tau(x) = \left[\frac{d}{d\nu}\mathbb{E}\big[Y_i(W_i + \nu) | X_i = x\big]\right]_{\nu=0}, \tag{9}$$

such that $V(\pi) = \mathbb{E}[\pi(X_i)\tau(X_i)]$ and regret $R(\pi)$ is as in (6). Our analysis pertains to any setup with a regret function $R(\pi)$ that admits such a representation. Given these preliminaries, recall that our goal is to learn low regret policies, that is, to use observational data to derive a policy $\hat{\pi} \in \Pi$ with a guarantee that $R(\hat{\pi}) = \mathcal{O}_P(1/\sqrt{n})$. In order to do so, we need to make assumptions on the observational data generation distribution that allow for identification and adequate estimation of $V(\pi)$, and also control the size of Π in a way that makes emulating the best-in-class policy a realistic objective. The following two subsections outline these required conditions; our main result is then stated in Section 2.3.

2.1. *Identifying and Estimating Causal Effects*

In order to learn a good policy $\hat{\pi}$, we first need to be able to evaluate $V(\pi)$ for any specific policy π. Our main assumption, following Chernozhukov et al. (2016), is that we can construct a doubly robust score for the average treatment effect $\theta = \mathbb{E}[\tau(X_i)]$. At the end of this section, we discuss how this approach applies to three important examples, and refer the reader to Chernozhukov et al. (2016) for a more general discussion of when such doubly robust scores exist.

ASSUMPTION 1: *Write $m(x, w) = \mathbb{E}[Y_i(w) | X_i = x] \in \mathcal{M}$ for the counterfactual response surface. We assume that $m(x, w)$ induces a treatment effect function $\tau_m(x, w)$ with the following properties*:

1. *The functional $m(\cdot) \to \tau_m(\cdot)$ is linear in m, and there exists a weighting function $g(x, z)$ that identifies $\tau_m(\cdot)$ via*

$$\mathbb{E}\big[\tau_{\tilde{m}}(X_i, W_i) - g(X_i, Z_i)\tilde{m}(X_i, W_i) | X_i\big] = 0, \tag{10}$$

   for any counterfactual response surface $\tilde{m}(x, w) \in \mathcal{M}$.
2. *Policy value can be defined in terms of moments of $\tau_m(X_i, W_i)$, such that $V(\pi) = \mathbb{E}[\pi(X_i)\tau(X_i)]$ with $\tau(x) = \mathbb{E}[\tau_m(X_i, W_i) | X_i = x]$ for all $\pi : \mathcal{X} \to \{0, 1\}$.*

In some examples, $\tau_m(x, w)$ does not depend on w, and we omit the w-argument of $\tau_m(\cdot)$.

[6] Another interesting question one could ask is how best to optimize the assignment of W_i globally rather than locally (i.e., the case where we can set the treatment level w to an arbitrary level, rather than simply nudge the preexisting levels of W_i). This question would require different formal tools, however, as the results developed in this paper only apply to binary decisions.

Given this setup, Chernozhukov et al. (2016) proposed first estimating $g(\cdot)$ and $m(\cdot)$, and then considered

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_i, \qquad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i, W_i) + \hat{g}(X_i, Z_i)\big(Y_i - \hat{m}(X_i, W_i)\big). \tag{11}$$

They show that this estimator is $\sqrt{n}$-consistent and asymptotically unbiased Gaussian for θ, provided that the nuisance estimates $\hat{g}(\cdot)$ and $\hat{m}(\cdot)$ converge sufficiently fast and that we use cross-fitting (Chernozhukov et al. (2018), Schick (1986)). This estimator is also semiparametrically efficient under general conditions (Newey (1994)).[7]

Our approach to policy learning builds on these foundations. We again start by estimating nuisance components and by forming doubly robust scores as in (11). However, instead of just averaging the $\widehat{\Gamma}_i$ to estimate θ, we use these scores for policy learning by plugging them into (2). Our main result will establish that we can get strong regret bounds for learning policies under conditions that are similar to those used by Chernozhukov et al. (2016) to show asymptotic normality of (11) and, more broadly, that build on assumptions often made in the literature on semiparametric efficiency (Bickel et al. (1998), Chen, Hong, and Tarozzi (2008), Hahn (1998), Newey (1994), Robins and Rotnitzky (1995)).

As in the recent work of Chernozhukov et al. (2018) on double machine learning or that of van der Laan and Rose (2011) on targeted learning, we take an agnostic view on how the nuisance estimates $\hat{g}(\cdot)$ and $\hat{m}(\cdot)$ are obtained, and simply impose high level conditions on their rates of convergence. Given sufficient regularity, we can construct estimators that satisfy the rate condition (13) via, for example, sieve-based methods (Chen (2007)) or kernel regression (Caponnetto and De Vito (2007)). Moreover, in applications, we may want to consider several different machine learning methods for each component, or potentially combinations thereof, and then use cross-validation to choose which method to use. For completeness, we allow problem specific quantities to change with the sample size n, and track this dependence with a subscript n, for example, $m_n(x, w) = \mathbb{E}_n[Y_i(w)|X_i = x]$, etc.

ASSUMPTION 2: *In the setting of Assumption 1, assume that second moments are controlled as* $\mathbb{E}_n[m_n^2(X_i, W_i)]$, $\mathbb{E}_n[\tau_{m_n}^2(X_i, W_i)] < \infty$ *and* $\mathbb{E}_n[g_n^2(X_i, Z_i)] < \infty$ *for all* $n = 1, 2, \ldots$, *and that we have access to uniformly consistent estimators of these nuisance components*,

$$\begin{aligned} \sup_{x,w}\big\{\big|\hat{m}_n(x, w) - m_n(x, w)\big|\big\}, \sup_{x,w}\big\{\big|\tau_{\hat{m}_n}(x, w) - \tau_{m_n}(x, w)\big|\big\} \to_p 0, \\ \sup_{x,z}\big\{\big|\hat{g}_n(x, z) - g_n(x, z)\big|\big\} \to_p 0, \end{aligned} \tag{12}$$

whose L_2 *errors decay as follows, for some* $0 < \zeta_m, \zeta_g < 1$ *with* $\zeta_m + \zeta_g \geq 1$ *and some* $a(n) \to 0$, *where* (X, W, Z) *is taken to be an independent test example drawn from the same*

[7] Our results do not depend on efficiency of (11); rather, we only use $\sqrt{n}$-consistency. In cases where (11) may not be efficient, our regret bounds still hold verbatim; the only difference being that we can no longer interpret the terms of the form $\mathbb{E}[\Gamma_i^2]$ appearing in the bound as related to the semiparametric efficient variance for θ.

distribution as the training data:[8]

$$\mathbb{E}\Big[\big(\hat{m}_n(X,W) - m_n(X,W)\big)^2\Big], \mathbb{E}\Big[\big(\tau_{\hat{m}_n}(X,W) - \tau_{m_n}(X,W)\big)^2\Big] \leq \frac{a(n)}{n^{\xi_m}},$$
$$\mathbb{E}\Big[\big(\hat{g}_n(X,Z) - g_n(X,W)\big)^2\Big] \leq \frac{a(n)}{n^{\xi_g}}. \tag{13}$$

We end this section by verifying that Assumption 1 in fact covers several settings of interest, and is closely related to several standard approaches to semiparametric inference. In cases with selection on observables, we do not need an instrument (or can simply set $Z_i = W_i$), so for simplicity of notation we replace all instances of Z_i with W_i.

Binary Treatment With Selection on Observables. Most existing work on policy learning, including Kitagawa and Tetenov (2018), has focused on the setup where W_i is binary and unconfounded, that is, $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i | X_i$. In this case, weighting by the inverse propensity score lets us recover the average treatment effect, that is, $g(x,w) = (w - e(x))/(e(x)(1-e(x)))$ with $e(x) = \mathbb{P}[W_i = 1 | X_i = x]$ identifies the conditional average treatment effect $\tau_m(x) = m(x,1) - m(x,0)$ via (10). The estimation strategy (11) yields

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{m}(X_i,1) - \hat{m}(X_i,0) + \frac{W_i - \hat{e}(X_i)}{\hat{e}(X_i)\big(1-\hat{e}(X_i)\big)}\big(Y_i - \hat{m}(X_i,W_i)\big)\Big), \tag{14}$$

and recovers augmented inverse propensity weighting (Robins, Rotnitzky, and Zhao (1994)).

Continuous Treatment With Selection on Observables. In the case where W_i is continuous and unconfounded $\{Y_i(w)\} \perp\!\!\!\perp W_i | X_i$, we can derive a representer $g(\cdot)$ via integration by parts (Powell, Stock, and Stoker (1989)). Under regularity conditions, the τ-function $\tau_m(x,w) = [d/d\nu m(x, w+\nu)]_{\nu=0}$ can be identified via (10) using

$$\int\int \frac{d}{dw}\big[m(X_i,W_i)\big]_{w=W} dF_{W_i|X_i}\, dF_{X_i} = \int\int g(X_i,W_i) m(X_i,W_i)\, dF_{W_i|X_i}\, dF_{X_i},$$
$$g(X_i,W_i) = -\frac{d}{dw}\big[\log\big(f(w|X_i)\big)\big]_{w=W_i}, \tag{15}$$

where $f(\cdot|x)$ denotes the conditional density of W_i given $X_i = x$. The resulting doubly robust estimator was to our knowledge first derived via the general approach of Chernozhukov et al. (2016), which in turn is closely related to an approach proposed by Ai and Chen (2007).

[8] A notable special case of this assumption is when $\zeta_m = \zeta_g = 1/2$; this is equivalent to the standard assumption in the semiparametric estimation literature that all nuisance components (i.e., in our case, both the outcome and weighting regressions) are $o(n^{-1/4})$-consistent in terms of L_2-error. The weaker requirement (13) reflects the fact that doubly robust treatment effect estimators can trade-off accuracy of the m-model with accuracy of the g-model, provided the product of the error rates is controlled (Farrell (2015)).

Binary, Endogenous Treatment With Binary Treatment and Instrument. Instead of unconfoundedness, now suppose that Z_i is a valid instrument conditionally on features X_i in the sense of Assumption 2.1 of Abadie (2003). Suppose moreover that treatment effects are homogenous, meaning that the conditional average treatment effect matches the conditional local average treatment effect (Imbens and Angrist (1994)),[9]

$$\tau_m(x) = m(x,1) - m(x,0) = \frac{\mathrm{Cov}[Y_i, Z_i | X_i = x]}{\mathrm{Cov}[W_i, Z_i | X_i = x]}. \tag{16}$$

Then we can use a weighting function $g(\cdot)$ defined in terms of the compliance score (Abadie (2003), Aronow and Carnegie (2013)),

$$\begin{aligned} g(X_i, Z_i) &= \frac{1}{\Delta(X_i)} \frac{Z_i - z(X_i)}{z(X_i)(1 - z(X_i)}, \\ z(x) &= \mathbb{P}[Z_i = 1 | X_i = x], \\ \Delta(x) &= \mathbb{P}[W_i = 1 | Z_i = 1, X_i = x] - \mathbb{P}[W_i = 1 | Z_i = 0, X_i = x], \end{aligned} \tag{17}$$

to identify this τ-function using (10). We note that our formal results all require that $g(\cdot)$ be bounded, which implicitly rules out the case of weak instruments (since if Δ approaches 0, the $g(\cdot)$-weights blow up).

2.2. *Assumptions About the Policy Class*

Next, in order to obtain regret bounds that decay as $1/\sqrt{n}$, we need some control over the complexity of the class Π (and again let Π potentially change with n for generality). The Vapnik–Chervonenkis (VC) approach (Vapnik (2000)) presents us with a natural way to do so. Recall that the VC-dimension of a class Π of binary decision rules is the largest value of $d \in \mathbb{N}$ such that there exists a set of d points $x_1, \ldots, x_d \in \mathcal{X}$ that is "shattered" by Π in the following sense: For each 2^d of the binary vectors $v \in \{0,1\}^d$, there exists a policy $\pi_v \in \Pi$ such that $\pi_v(X_i) = v_i$ for all $i = 1, \ldots, d$. Throughout our analysis, we control the complexity of Π_n by assuming that its VC-dimension does not grow too fast with the sample size n. As is familiar from the literature on classification, we will find that the best possible uniform regret bounds scale as $\sqrt{\mathrm{VC}(\Pi_n)/n}$ (Vapnik (2000)).

ASSUMPTION 3: *We assume that there are constants* $0 < \beta < 1/2$ *and* $N \geq 1$ *such that the Vapnik–Chervonenkis dimension of* Π_n *is bounded as* $\mathrm{VC}(\Pi_n) \leq n^\beta$ *for all* $n \geq N$.

In order to illustrate this assumption, we give two examples of policy classes that have a finite VC dimension, and one that does not. In all three examples below, we assume that the features X_i take values in $\mathcal{X} = \mathbb{R}^p$ for some $p \geq 1$.

Linear Rules. The VC-dimension of the class of linear decision rules is (Wainwright (2019, p. 116)) $\mathrm{VC}(\Pi) = p + 1$ for $\Pi = \{\pi_{v,c} : \pi_{v,c}(x) = 1(\{v \cdot x \geq c\}), v \in \mathbb{R}^p, c \in \mathbb{R}\}$. Thus, our approach applies to linear decision rules in dimension $p_n \leq n^\beta$ for some $\beta < 1/2$.

[9] As discussed above, our notation has potential outcomes $Y_i(W_i)$ that only depend on treatment W_i, and do not involve the instrument Z_i. This is only meaningful when the exclusion restriction holds.

Decision Trees. Trees represent decision rules recursively (Breiman, Friedman, Olshen, and Stone (1984)). A depth-0 decision tree T_0 is a trivial decision rule, $T_0(x) = a$ for some $a \in \{0, 1\}$ and all $x \in \mathcal{X}$. For any $L \geq 1$, a depth-L decision tree T_L is specified via a splitting variable $j \in 1, \ldots, p$, a threshold $t \in \mathbb{R}$, and two depth-$(L-1)$ decision trees $T_{(L-1),A}$ and $T_{(L-1),B}$, such that $T_L(x) = T_{(L-1),A}(x)$ if $x_j \leq t$, and $T(x) = T_{(L-1),B}(x)$ else. See Figure 1 for an example of a decision tree. The class of depth-L decision trees over $\mathbb{R}^p$ has VC dimension bounded on the order of $\text{VC}(\Pi) = \widetilde{\mathcal{O}}(2^L \log(p))$.[10] Thus, our results apply to trees whose depth may grow as $L_n = \lfloor \kappa \log_2(n) \rfloor$ for some $\kappa < 1/2$.

Monotone Rules. We have $x \in [0, 1]^2$ and units get treated if x_2 exceeds some increasing function of x_1, that is, $\Pi = \{\pi_f : \pi_f(x) = 1(\{x_2 \geq f(x_1)\}), f \text{ is monotone increasing}\}$. This class has infinite VC dimension, because any set of points $\{x_i\}_{i=1}^d$ with $x_i = (\alpha_i, \alpha_i^2)$ and $0 < \alpha_1 < \cdots < \alpha_d < 1$ can be shattered using Π. Thus, our results do not apply to monotone rules over $[0, 1]^2$.[11]

2.3. *Bounding Asymptotic Regret*

We are now ready to state our main result on the asymptotic regret of policy learning using doubly robust scores. Following Chernozhukov et al. (2018, 2016), we assume that we run our method with scores obtained via cross-fitting, which is a type of data splitting that can be used to verify asymptotic normality given only high-level conditions on the predictive accuracy of the methods used to estimate nuisance components. In particular, cross-fitting allows for the use of black-box machine learning tools provided we can verify that they are accurate in mean-squared error as in Assumption 2.

We proceed as follows: First, divide the data into K evenly-sized folds and, for each fold $k = 1, \ldots, K$, run an estimator of our choice on the other $K - 1$ data folds to estimate the functions $m_n(x, w)$ and $g_n(x, z)$; denote the resulting estimates $\hat{m}_n^{(-k)}(x, w)$ and $\hat{g}_n^{(-k)}(x, z)$. Throughout, we will only assume that these nuisance estimates are accurate in the sense of Assumption 2. Then, given these precomputed values, we choose $\hat{\pi}_n$ by maximizing a doubly robust estimate of $A(\pi) = 2V(\pi) - \mathbb{E}[\tau(X_i)]$,

$$\begin{aligned} \hat{\pi}_n &= \text{argmax}\big\{\widehat{A}_n(\pi) : \pi \in \Pi_n\big\}, \\ \widehat{A}_n(\pi) &= \frac{1}{n}\sum_{i=1}^{n}\big(2\pi(X_i) - 1\big)\widehat{\Gamma}_i, \\ \widehat{\Gamma}_i &= \tau_{\hat{m}_n^{(-k(i))}}(X_i, W_i) + \hat{g}_n^{(-k(i))}(X_i, Z_i)\big(Y_i - \hat{m}_n^{(-k(i))}(X_i, W_i)\big), \end{aligned} \tag{18}$$

[10] This bound follows Lemma 4 of Zhou, Athey, and Wager (2018), paired with the alternative characterization of the VC dimension given in Section A of the Supplemental Material (Athey and Wager (2021)). The notation $f(n) = \widetilde{\mathcal{O}}(g(n))$ means that there is a function $h(\cdot)$ that scales polylogarithmically in its argument for which $f(n) \leq h(g(n))g(n)$.

[11] The difficulty here is not a mere technicality: Monotone decision rules can match arbitrary decision rules along the curve (α, α^2) for $\alpha \in [0, 1]$, and so it is impossible to establish any nontrivial learning rates over monotone decision rules without making further assumptions on the distribution of the features X_i. In particular, we need assumptions that guarantee that all observations cannot concentrate around the curve (α, α^2). In this paper, we do not consider results that require specific distributional assumptions over the features X_i. We note however the recent work by Mbakop and Tabord-Meehan (2016), who establish polynomial rates of convergence for learning monotone rules under an assumption that the X_i have a bounded density under Lebesgue measure on $[0, 1]^2$.

where $k(i) \in \{1, \ldots, K\}$ denotes the fold containing the ith observation. The K-fold algorithmic structure used in (18) was proposed in an early paper by Schick (1986) as a general purpose tool for efficient estimation in semiparametric models, and has also been used by other authors including Robins, Li, Mukherjee, Tchetgen, and van der Vaart (2017) and Zheng and van der Laan (2011).

Finally, we assume that the weighting function $g_n(x, z)$ is bounded uniformly as below. In the case of a binary exogenous treatment, this is equivalent to the "overlap" assumption in the causal inference literature (Imbens and Rubin (2015)), whereby $\eta \leq \mathbb{P}[W_i = 1 | X_i = x] \leq 1 - \eta$ for all values of x. In our setting, the condition below acts as a generalization of the overlap assumption (Hirshberg and Wager (2018)).

ASSUMPTION 4: *There is an* $\eta > 0$ *such that* $|g_n(x, z)| \leq \eta^{-1}$ *for all* x, z, n.

We also define the following quantities, where S_n bounds the second moment of the scores, and S_n^* is the asymptotic variance for estimating the policy improvement $A(\pi)$ of the best policy in Π_n via (11):[12]

$$\begin{aligned} S_n &= \mathbb{E}\Big[\big(\tau_{m_n}(X_i, W_i) - g_n(X_i, Z_i)\big(Y_i - m_n(X_i, W_i)\big)\big)^2\Big], \\ S_n^* &= \inf\Big\{\mathrm{Var}\Big[\big(2\pi(X_i) - 1\big)\big(\tau_{m_n}(X_i, W_i) - g_n(X_i, Z_i)\big(Y_i - m_n(X_i, W_i)\big)\big)\Big] \\ &\qquad : \pi \in \Pi_n\Big\}. \end{aligned} \tag{19}$$

We note that, unless we have an exceptionally large signal-to-noise ratio, we will have $S_n^* \geq S_n/4$ and so the rounded log-term in (20) below is just 0. A proof of Theorem 1 is given in the following section.

THEOREM 1: *Given Assumptions 1, 2, and 4, define* $\hat{\pi}_n$ *as in* (18).[13] *Suppose moreover that the irreducible noise* $\varepsilon_i = Y_i - m(X_i, W_i)$ *is both uniformly sub-Gaussian conditionally on* X_i *and* W_i *and has second moments uniformly bounded from below,* $\mathrm{Var}[\varepsilon_i | X_i = x, W_i = w] \geq s^2$, *and that the treatment effect function* $\tau_{m_n}(x, w)$ *is uniformly bounded in* x, w, *and* n. *Finally, suppose that* Π_n *satisfies Assumption 3 with parameter* $\beta \leq \min\{\zeta_m, \zeta_g\}$, *where the* ζ *are as defined in Assumption 2. Then, for any sequence* $\psi_n \geq 0$ *with* $\lim_{n\to\infty} \psi_n \sqrt{n} = 0$,

$$\begin{aligned} &\limsup_{n\to\infty} \mathbb{E}\Big[\sup\big\{R_n(\pi) : \widehat{A}_n(\pi) \geq \max\big\{\widehat{A}_n(\pi) : \pi \in \Pi_n\big\} - \psi_n, \pi \in \Pi_n\big\}\Big] \\ &\Big/ \sqrt{\mathrm{VC}(\Pi_n) S_n^* \left(1 + \left\lfloor \log_4\left(\frac{S_n}{S_n^*}\right)\right\rfloor / 9\right) / n} \leq 60, \end{aligned} \tag{20}$$

where $R_n(\cdot)$ *denotes regret for the nth data-generating distribution*.

In the simplest case where the maximizer of $\widehat{A}_n(\pi)$ over $\pi \in \Pi_n$ is unique and $\psi_n = 0$ (i.e., we solve the maximization problem exactly), the statement in (20) simplifies to a

[12] By expanding the square, we see that policies with higher values have lower variance of their scores, and so S_n^* corresponds to the asymptotic variance for evaluating an optimal policy. Moreover, in the case where arguments from Newey (1994) imply that the doubly robust estimator (11) is efficient, then S_n^* is the semiparametric efficient variance for evaluating an optimal policy.

[13] We assume that the rates of convergence specified in Assumption 2 apply to the nuisance components estimated for each fold $k = 1, \ldots, K$ in (18).

bound on $\mathbb{E}[R_n(\hat{\pi}_n)]$, where $\hat{\pi}_n$ is as defined in (18). However, in practice, $\widehat{A}_n(\pi)$ may have many maximizers. Moreover, the optimization problem (18) is not convex and so—given a reasonable computational budget—we may only be able to solve it to within some tolerance $\psi_n > 0$. The more comprehensive form of our result given above highlights the fact that, in this case, our regret bound in fact applies uniformly over all approximate solutions to (18).

3. UPPER BOUNDS

In this section, we present a series of results that culminate in a proof of Theorem 1, given in Section 3.3. All other proofs are deferred to Section C of the Online Supplemental Material (Athey and Wager (2021)). Recall that we study policy learning for a class of problems where regret can be written as in (6) using a function $V_n(\pi) = \mathbb{E}_n[\pi(X_i)\tau_n(X_i)]$, and we obtain $\hat{\pi}_n$ by maximizing a cross-fitted doubly robust estimate of $A_n(\pi) = 2V_n(\pi) - \mathbb{E}_n[\tau_n(X_i)]$ defined in (18) over the class Π_n. If we could use $\widehat{A}_n(\pi) = A_n(\pi)$, then (18) would directly yield the regret-minimizing policy in the class Π_n; but of course we never know $A_n(\pi)$ in applications. Thus, the main focus of our formal results is to study stochastic fluctuations of the empirical process $\widehat{A}_n(\pi) - A_n(\pi)$ for $\pi \in \Pi_n$, and examine how they affect the quality of policies learned via (18).

3.1. *Rademacher Complexities and Oracle Regret Bounds*

We start our analysis by characterizing concentration of an ideal version of the objective in (18) based on the true influence scores Γ_i, rather than doubly robust estimates thereof:

$$\widetilde{A}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n}\big(2\pi(X_i) - 1\big)\Gamma_i,$$
$$\Gamma_i = \tau_{m_n}(X_i, W_i) + g_n(X_i, Z_i)\big(Y_i - m_n(X_i, W_i)\big). \tag{21}$$

The advantage of studying concentration of the empirical process $\widetilde{A}_n(\pi) - A_n(\pi)$ over the set $\pi \in \Pi_n$ is that it allows us, for the time being, to abstract away from the estimation tools used to obtain $\widehat{A}_n(\pi)$, and instead to focus on the complexity of empirical maximization over the class Π_n.

A convenient way to bound the supremum of this empirical process over any class Π is by controlling its Rademacher complexity $\mathcal{R}_n(\Pi)$, defined as[14]

$$\mathcal{R}_n(\Pi) = \mathbb{E}\left[\sup_{\pi\in\Pi}\left\{\frac{1}{n}\sum_{i=1}^{n}\xi_i\Gamma_i\big(2\pi(X_i) - 1\big)\right\}\Big|\{X_i, \Gamma_i\}_{i=1}^n\right] \tag{22}$$

where the ξ_i are independent Rademacher (i.e., sign) random variables $\xi_i = \pm 1$ with probability 1/2 each (Bartlett and Mendelson (2002)). For intuition as to why Rademacher complexity is a natural complexity measure, note that $\mathcal{R}_n(\Pi)$ characterizes the maximum (weighted) in-sample classification accuracy on randomly generated labels ξ_i over classifiers $\pi \in \Pi$; thus, $\mathcal{R}_n(\Pi)$ measures how much we can overfit to random coin flips using Π.

[14]Note that, conditionally on $\{X_i, \Gamma_i\}_{i=1}^n$ and the Rademacher variables ξ_i, the sum $\sum_{i=1}^n \xi_i\Gamma_i(2\pi(X_i) - 1)$ can only take 2^n distinct values. Thus, the definition of $\mathcal{R}_n(\Pi)$ does not entail any measure theoretic problems.

Following this proof strategy, we bound the Rademacher complexity of "slices" of our policy class Π_n, defined as

$$\Pi_n^\lambda = \{\pi \in \Pi_n : R_n(\pi) \leq \lambda\}. \tag{23}$$

The reason we focus on slices of Π_n is that, when we use doubly robust scores, low-regret policies can generally be evaluated more accurately than high-regret policies, and using this fact allows for sharper bounds. Specifically, we can check that $n\text{Var}[\widetilde{A}_n(\pi)] = S_n - A_n^2(\pi)$, and so

$$n \sup\{\text{Var}[\widetilde{A}_n(\pi)] : \pi \in \Pi_n^\lambda\} := S_n^\lambda \leq S_n^* + 4\lambda \sup\{A_n(\pi) : \pi \in \Pi_n\}, \tag{24}$$

where S_n and S_n^* are defined in (19). This type of slicing technique is common in the literature, and has been used in different contexts by, for example, Bartlett, Bousquet, and Mendelson (2005) and Giné and Koltchinskii (2006).

The following result provides such a bound in terms of the second moments of the doubly robust score, specifically S_n^λ and S_n. This bound is substantially stronger than corresponding bounds used in existing results on policy learning. Kitagawa and Tetenov (2018) built their result on bounds that depend on $\max\{\Gamma_i\}/\sqrt{n}$, which can only be used with scores that are uniformly bounded in order to get optimal rates. Meanwhile, bounds that scale as $\sqrt{S_n^\lambda \log(n)/n}$ are developed by Cortes, Mansour, and Mohri (2010), Maurer and Pontil (2009), and Swaminathan and Joachims (2015); however, the additional $\log(n)$ factor makes these bounds inappropriate for asymptotic analysis.

LEMMA 2: *Suppose that the class Π_n satisfies Assumption 3, and that the scores Γ_i in (21) are drawn from a sequence of uniformly sub-Gaussian distributions with variance bounded from below*:

$$\mathbb{P}_n[|\Gamma_i| > t] \leq C_\nu e^{-\nu t^2} \quad \textit{for all } t > 0, \qquad \text{Var}_n[\Gamma_i | X_i = x] \geq s^2, \tag{25}$$

for some constants $C_\nu, \nu, s > 0$ and all $n = 1, 2, \ldots$ Then, for any λ,

$$\limsup_{n\to\infty} \mathbb{E}[\mathcal{R}_n(\Pi_n^\lambda)] \Big/ \sqrt{(S_n^\lambda + 4\lambda^2)\left(1 + \left\lfloor \log_4\left(\frac{S_n}{S_n^\lambda}\right)\right\rfloor/9\right)\frac{\text{VC}(\Pi_n)}{n}} \leq 20. \tag{26}$$

Then, following the well-known approach of Bartlett and Mendelson (2002), we use our bound on Rademacher complexity to obtain a uniform concentration bound for $\widetilde{A}_n(\pi)$. We use a refinement of the argument of Bartlett and Mendelson (2002) based on Talagrand's inequality to get a bound that depends on second moments of Γ_i rather than $\sup|\Gamma_i|$.

COROLLARY 3: *Under the conditions of Lemma 2, the expected maximum error of $\widetilde{A}_n(\pi)$ is bounded as*

$$\limsup_{n\to\infty} \mathbb{E}\left[\sup\left\{\left|\widetilde{A}_n(\pi) - A_n(\pi)\right| : \pi \in \Pi_n^\lambda\right\}\right] \Big/ \sqrt{(S_n^\lambda + 4\lambda^2)\left(1 + \left\lfloor \log_4\left(\frac{S_n}{S_n^\lambda}\right)\right\rfloor/9\right)\frac{\text{VC}(\Pi_n)}{n}} \leq 40. \tag{27}$$

Furthermore, this error is concentrated around its expectation: There is a sequence $c_n \to 0$ *such that, for any* $\delta > 0$,

$$\sup\left\{\left|\widetilde{A}_n(\pi) - A_n(\pi)\right| : \pi \in \Pi_n^\lambda\right\} \leq (1 + c_n)\left(\mathbb{E}\left[\sup\left\{\left|\widetilde{A}_n(\pi) - A_n(\pi)\right| : \pi \in \Pi_n^\lambda\right\}\right] + \sqrt{\frac{2S_n^\lambda \log\left(\delta^{-1}\right)}{n}}\right) \tag{28}$$

with probability at least $1 - \delta$.

In our final argument, we will apply Corollary 3 for different λ-slices, and verify that we can in fact focus on those slices where λ is nearly 0. Before that, however, we also need to control the discrepancy between the feasible objective $\widehat{A}_n(\pi)$ and the oracle surrogate $\widetilde{A}_n(\pi)$ studied here.

3.2. *Uniform Coupling With the Doubly Robust Estimator*

In the previous section, we established risk bounds that would hold if we could optimize the infeasible value function $\widetilde{A}_n(\pi)$; we next need to extend these bounds to cover the situation where we optimize a feasible value function. As discussed above, we focus on the doubly robust estimator (18), obtained using cross-fitting as in Chernozhukov et al. (2018, 2016). As preliminaries, we note that the results of Chernozhukov et al. (2016) immediately imply that, given Assumption 2, $\widehat{A}_n(1)$ is an asymptotically normal estimate of $A_n(1)$, where we use "1" as shorthand for the "always treat" policy. Furthermore, it is easy to check that given any fixed policy π,

$$\sqrt{n}\left(\widehat{A}_n(\pi) - \widetilde{A}_n(\pi)\right) \to_p 0, \tag{29}$$

meaning that the discrepancy between the two value estimates decays faster than the variance of either.

However, in our setting, the analyst gets to optimize over all policies $\pi \in \Pi_n$, and so coupling results established for a single predetermined policy π are not strong enough. The following lemma extends the work of Chernozhukov et al. (2016) to the case where we seek to establish a coupling of the form (29) that holds simultaneously for all $\pi \in \Pi_n$.

LEMMA 4: *Under the conditions of Lemma 2, suppose that Assumptions 1 and 4 hold, and that we obtain* $\widehat{A}_n(\pi)$ *using cross-fitted estimates of nuisance components satisfying Assumption 2. Then*

$$\frac{\sqrt{n}\mathbb{E}\left[\sup\left\{\left|\widehat{A}_n(\pi) - \widetilde{A}_n(\pi)\right| : \pi \in \Pi_n\right\}\right]}{a\left(\left(1 - K^{-1}\right)n\right)} = \mathcal{O}\left(1 + \sqrt{\frac{\mathrm{VC}(\Pi_n)}{n^{\min\{\zeta_m, \zeta_g\}}}}\right), \tag{30}$$

where the $\mathcal{O}(\cdot)$ *term hides a dependence on the overlap parameter* η *from Assumption 4 and the sub-Gaussianity parameter* ν *specified in Lemma 2.*

The above result is perhaps surprisingly strong: Provided that the dimension $\mathrm{VC}(\Pi_n)$ of Π_n does not grow too fast with n, the bound (30) is the same coupling bound as we might expect to obtain for a single policy π, and the dimension of the class Π_n does not affect the leading-order constants in the bound. In other words, in terms of the coupling

of $\widetilde{A}_n(\pi)$ and $\widehat{A}_n(\pi)$, we do not lose anything by scanning over a continuum of policies $\pi \in \Pi_n$ rather than just considering a single policy π.

The doubly robust form used here is not the only way to construct efficient estimators for the value of a single policy π, for example, Hirano, Imbens, and Ridder (2003) showed that inverse-propensity weighting with nonparametrically estimated propensity scores may also be efficient—but it plays a key role in the proof of Lemma 4. In particular, under Assumption 2, the natural bound for the bias term due to misspecification of the nuisance components in fact holds simultaneously for all $\pi \in \Pi$, and this helps us pay a smaller-than-expected price for seeking a uniform result as in (30). It is far from obvious that other efficient methods for evaluating a single policy π, such as that of Hirano, Imbens, and Ridder (2003), would lead to equally strong uniform couplings over the whole class Π_n.

3.3. *Proof of Theorem 1*

Given that Assumptions 1, 2, 3, and 4 hold with parameters $\beta < \min\{\zeta_m, \zeta_g\}$, a combination of results from Corollary 3 and Lemma 4 implies that $\widehat{A}_n(\cdot)$ concentrates around $A_n(\cdot)$ over Π_n^λ. To conclude, it now remains to apply these bounds at two different values of λ. First, we choose $\lambda^* > 0$ such as to satisfy $4(\lambda^*)^2 + 4\lambda^* \sup\{A(\pi) : \pi \in \Pi_n\} \le S_n^*$, so that the following holds via (24):

$$S_n^{\lambda^*} + 4\big(\lambda^*\big)^2 \le S_n^* + 4\big(\lambda^*\big)^2 + 4\lambda^* \sup\big\{A(\pi) : \pi \in \Pi_n\big\} \le 2S_n^*.$$

Then, by Corollary 3 and Lemma 4, we find that the limsup of the following expression is bounded by 1 as n goes to infinity:

$$\mathbb{E}\big[\sup\big\{\big|\widehat{A}_n(\pi) - A_n(\pi)\big| : \pi \in \Pi_n^{\lambda^*}\big\}\big] \Big/ \left(60\sqrt{S_n^*\left(1 + \left\lfloor \log_4\left(\frac{S_n}{S_n^*}\right)\right\rfloor/9\right)\frac{\mathrm{VC}(\Pi_n)}{n}}\right).$$

Now, recall that if any two functions $h(\cdot)$ and $\hat{h}(\cdot)$ are uniformly coupled as $|h(u) - \hat{h}(u)| \le b$ for all $u \in U$ and $\hat{h}(\hat{u}) \ge \sup\{\hat{h}(u) : u \in U\} - \psi$, then

$$h(\hat{u}) \ge \hat{h}(\hat{u}) - b \ge \hat{h}(u) - b - \psi \ge h(u) - 2b - \psi$$

for any $u \in U$. Thus, the above implies that (recall that $A_n(\pi)$ scales with $2R_n(\pi)$)

$$\limsup_{n\to\infty} \mathbb{E}\big[\sup\big\{R_n(\pi) : \widehat{A}_n(\pi) \ge \max\big\{\widehat{A}_n(\pi) : \pi \in \Pi_n^{\lambda^*}\big\} - \psi_n, \pi \in \Pi_n^{\lambda^*}\big\}\big]$$
$$\Big/ \left(\frac{\psi_n}{2} + 60\sqrt{S_n^*\left(1 + \left\lfloor \log_4\left(\frac{S_n}{S_n^*}\right)\right\rfloor/9\right)\frac{\mathrm{VC}(\Pi_n)}{n}}\right) \le 1, \tag{31}$$

and we note that ψ_n decays fast enough by assumption that it can be omitted from (31) without altering the result. In other words, if we knew that our learned policy approximately maximizes $\widehat{A}_n(\pi)$ *and* has regret less than λ^*, then we could guarantee that its regret decays at the desired rate.

To prove our result, it remains to show that all approximate maximizers of $\widehat{A}_n(\cdot)$ have regret bounded by λ^* enough for (31) to capture the leading-order behavior of regret.

To do so, we apply a similar argument as above, but at a different value of λ. Consider $\lambda_+ = 3\limsup_{n\to\infty} \sup\{R_n(\pi) : \pi \in \Pi_n\}$, and by (28) we see that

$$\lim_{n\to\infty} \sqrt{n}\mathbb{P}\left[\sup\left\{\left|\widetilde{A}_n(\pi) - A_n(\pi)\right| : \pi \in \Pi_n^{\lambda_+}\right\} \geq \frac{\lambda^*}{5}\right] = 0. \tag{32}$$

Furthermore, note that $\Pi_n^{\lambda_+} = \Pi_n$ for large enough n, and so (32) in fact also holds with $\Pi_n^{\lambda_+}$ replaced by Π_n. Meanwhile, from (30) paired with Markov's inequality we know that

$$\mathbb{P}\left[\sup\left\{\left|\widehat{A}_n(\pi) - \widetilde{A}_n(\pi)\right| : \pi \in \Pi_n\right\} \geq \frac{\lambda^*}{5}\right] = \mathcal{O}\left(\frac{a\left(\left(1-K^{-1}\right)n\right)}{\sqrt{n}}\right). \tag{33}$$

By combining these two bounds, we see that

$$\begin{aligned}\lim_{n\to\infty} \sqrt{n}\mathbb{P}\Big[&\left\{\pi \in \Pi_n : \widehat{A}_n(\pi) \geq \max\left\{\widehat{A}_n(\pi) : \pi \in \Pi_n\right\} - \psi_n\right\} \\ &\cap \left\{\pi \in \Pi_n : R_n(\pi) \geq \lambda^*\right\} \neq \emptyset\Big] = 0,\end{aligned} \tag{34}$$

and moreover, because $\tau_{m_n}(x, w)$ is uniformly bounded, we find that the contribution of events where (34) fails to hold to (20) is vanishingly small as n gets large.

4. LOWER BOUNDS

To complement the upper bounds given in Theorem 1, we also present lower bounds on the minimax risk for policy learning. Our goal is to show that our bounds are the best possible regret bounds that flexibly account for the distribution of the observed data and depend on the policy class Π through the Vapnik–Chervonenkis dimension $\mathrm{VC}(\Pi)$. For simplicity, we here only consider the case where W_i is binary and unconfounded; lower bounds for other cases considered in this paper can be derived via analogous arguments.

To establish our result, we consider lower bounds over sequences of problems defined as follows. Let $\mathcal{X}_s := [0, 1]^s$ denote the s-dimensional unit cube for some positive integer s, and let $f(x)$ and $e(x)$ be $\lceil s/2 + 1 \rceil$ times continuously differentiable functions over $\mathcal{X}_s$. Moreover, let $\sigma^2(x)$ and $\tau(x)$ be functions on $\mathcal{X}_s$ such that $\sigma^2(x)$ is bounded away from 0 and ∞, and $|\tau(x)|$ is bounded away from ∞. Then we define an asymptotically ambiguous problem sequence as one where $\{X_i, Y_i, W_i\}$ are independently and identically distributed drawn as

$$\begin{aligned}X_i &\sim \mathcal{P},\ W_i|X_i \sim \mathrm{Bernoulli}\big(e(X_i)\big), \\ Y_i|X_i, W_i &\sim \mathcal{N}\left(f(X_i) + \big(W_i - e(X_i)\big)\frac{\tau(X_i)}{\sqrt{n}},\ \sigma^2(X_i)\right).\end{aligned} \tag{35}$$

Because of the number of derivatives assumed on $f(x)$ and $e(x)$, it is well known that simple series estimators satisfy Assumption 2.[15] Thus, because the magnitude of the treat-

[15] See Nickl and Pötscher (2007) for an argument that holds for arbitrary distributions $\mathcal{P}$ supported on $[0, 1]^s$. We also note that, for a complete argument, one needs to address the fact that we have not assumed the treatment effect function $\tau(x)$ to be differentiable. To address this issue, note that in our data-generating process (35) we have $\mathbb{E}[Y_i|X_i = x] = f(x)$ regardless of n. Thus, because both $e(x)$ and $f(x)$ are sufficiently differentiable, we can use standard results about series estimation to obtain $o_P(n^{-1/4})$-consistent estimators $\hat{e}(x)$ and $\hat{f}(x)$ for these quantities. Next, for the purpose of our policy learner, we simply set $\hat{m}(x, 0) = \hat{m}(x, 1) = \hat{f}(x)$; and because $\mathbb{E}[\tau^2(X_i)/\sqrt{n}] = \mathcal{O}(1/n)$, these regression adjustments in fact satisfy Assumption 2.

ment effects shrinks in (35), S_n^* and S_n both converge to $S_{\mathcal{P}}$ as defined below, and so Theorem 1 immediately implies that, under unconfoundedness,

$$\limsup_{n\to\infty} R_n(\hat{\pi}_n)/\sqrt{\frac{S_{\mathcal{P}}\,\mathrm{VC}(\Pi)}{n}} \le 60, \qquad S_{\mathcal{P}} = \mathbb{E}_{\mathcal{P}}\left[\frac{\sigma^2(X_i)}{e(X_i)\big(1-e(X_i)\big)}\right] \tag{36}$$

for any policy class Π with finite VC dimension. The following result shows that (36) is sharp up to a universal constant (whose value is less than 200).[16]

THEOREM 5: *Let $f(x)$, $e(x)$, and $\sigma(x)$ be functions over $\mathcal{X}_s$ satisfying the conditions discussed above, and let Π be a class of functions over $\mathcal{X}_s$ with finite VC dimension. Then there exists a distribution $\mathcal{P}$ supported on $[0,1]^s$ (and a constant C) such that the minimax risk for policy learning over the data generating distribution* (35) *(with unknown $|\tau(x)| \le C$) and the policy class Π is bounded from below as follows, where $\hat{\pi}_n$ can be any measurable function of the training sample*:

$$\liminf_{n\to\infty}\left\{\sqrt{n}\inf_{\hat{\pi}_n}\left\{\sup_{|\tau(x)|\le C}\left\{\mathbb{E}\big[R_n(\hat{\pi}_n)\big]\right\}\right\}\right\} \ge 0.33\sqrt{S_{\mathcal{P}}\,\mathrm{VC}(\Pi)}. \tag{37}$$

Here, the fact that we focus on problems where the magnitude of the treatment effect scales as $1/\sqrt{n}$ is important, and closely mirrors the type of asymptotics used by Hirano and Porter (2009). If treatment effects decay faster than $1/\sqrt{n}$, then learning better-than-random policies is effectively impossible—but this does not matter, because of course all decision rules have regret decaying as $o(1/\sqrt{n})$ and so Theorem 1 is loose. Conversely, if treatment effects dominate the $1/\sqrt{n}$ scale, then in large samples it is all but obvious who should be treated and who should not, and it is possible to get regret bounds that decay at superefficient rates (Luedtke and Chambaz (2020)), again making Theorem 1 loose. But if the treatment effects obey the $\Theta(1/\sqrt{n})$ scaling of Hirano and Porter (2009), then the problem of learning good policies is neither trivial nor impossible, and the value of using doubly robust policy evaluation for policy learning becomes apparent.

Finally, we note that the bounds of Kitagawa and Tetenov (2018) for inverse-propensity weighting are not asymptotically sharp in the above sense. Even when propensity scores are known, Kitagawa and Tetenov (2018) assumed that $|Y_i| \le M$ and $\eta \le e(X_i) \le 1-\eta$, and then prove regret bounds that scale as $M/\eta\sqrt{\mathrm{VC}(\Pi)/n}$ instead of $\sqrt{S_{\mathcal{P}}\,\mathrm{VC}(\Pi)/n}$ in (36). Now, the bound of Kitagawa and Tetenov (2018) is of course sometimes sharp, for example, it is optimal if all we know is that $|Y_i| \le M$ and $\eta \le e(X_i) \le 1-\eta$, but it is not adaptively sharp for asymptotically ambiguous sequences of problems as in (35). In particular, the ratio of the upper bound of Kitagawa and Tetenov (2018) and the lower bound (37) scales as $M/(\eta\sqrt{S_{\mathcal{P}}})$, and there exist sequences of type (35) where this ratio may be arbitrarily large.[17]

[16]The strategy of proving lower bounds relative to an adversarial feature distribution $\mathcal{P}$ is standard in the machine learning literature; see, for example, Devroye and Lugosi (1995). If we fix the distribution $\mathcal{P}$ a priori, then regret bounds for empirical risk minimization over Π based on structural summaries of Π (such as the VC dimension) may be loose (Bartlett and Mendelson (2006)); however, it is not clear how to exploit this fact other than by conducting ad-hoc analyses for specific choices of Π.

[17]Using the techniques developed in this paper, we can sharpen the bounds of Kitagawa and Tetenov (2018) and asymptotically replace M/η by $\mathbb{E}[Y_i^2/(e(X_i)(1-e(X_i)))]^{1/2}$. However, even this improved bound may exceed (37) by an arbitrarily large factor.

5. IMPLEMENTATION AND EXPERIMENTS

We now illustrate the value of doubly robust scoring techniques for policy learning using both an example from program evaluation and simulation studies. In Section 5.1, we revisit a randomized evaluation of California's GAIN program, while Section 5.2 presents a simulation study with endogenous treatment assignment. We present additional simulation results on nudge interventions to a continuous treatment variable in Section B of the Online Supplemental Material.

Recall that our approach to policy involves a 3-step algorithm. We start with a set of n independent and identically distributed training examples (X_i, Y_i, W_i, Z_i) and a class Π of acceptable policies. Then we:

1. Estimate the nuisance components $m(x, w)$ and $g(x, z)$ defined in Section 2.1,
2. Form doubly robust scores[18] $\widehat{\Gamma}_i = \tau_{\hat{m}}(X_i, W_i) + \hat{g}(X_i, Z_i)(Y_i - \hat{m}(X_i, W_i))$, with cross-fitting as discussed in Section 2.3, and
3. Select $\hat{\pi} \in \text{argmax}\{\sum_{i=1}^{n}(2\pi(X_i) - 1)\widehat{\Gamma}_i : \pi \in \Pi\}$.

The main points of freedom left to the analysts involve the choice of estimator for $m(\cdot)$ and $g(\cdot)$ in Step 1, and the implementation of the optimization problem in Step 3. We emphasize that the choice of estimator for $m(x, w)$ and $g(x, z)$ in Step 1 and the choice of policy class Π along with the optimizer used in Step 3 can be made fully independently.

For Theorem 1 to apply, the main requirement on the method used to estimate $m(x, w)$ and $g(x, z)$ in Step 1 is that its error decays fast enough in mean-squared error, as detailed in Assumption 2. Here, one option is to use nonparametric estimators for which we can precisely spell out when they satisfy Assumption 2, such as sieve-based methods (Chen (2007)) or kernel regression (Caponnetto and De Vito (2007)); another is to use more heuristic methods from the statistical learning literature, such as boosting, random forests, or neural networks, in the hope that they will empirically be more accurate in finite samples than sieve or kernel-based methods.[19] One possible compromise is to run both classical methods known to satisfy Assumption 2 asymptotically and heuristic statistical learning tools, and then synthesize the output of all models via cross-validation. As argued in van der Laan, Polley, and Hubbard (2007), this approach essentially matches the finite-sample accuracy of the best method under consideration while preserving the asymptotic guarantees of the classical ones.

Meanwhile, the optimization problem in Step 3 is not a convex optimization problem, and so solving it can be computationally challenging. Several authors, including Beygelzimer and Langford (2009), Kitagawa and Tetenov (2018), Zhang et al. (2012) and Zhao et al. (2012), have noted that this optimization problem is numerically equivalent to a weighted classification problem,

$$\hat{\pi} = \underset{\pi \in \Pi}{\text{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \lambda_i H_i \big(2\pi(X_i) - 1\big) \right\}, \qquad \lambda_i = |\widehat{\Gamma}_i|, \qquad H_i = \text{sign}(\widehat{\Gamma}_i), \tag{38}$$

where we train a classifier $\pi(\cdot)$ with response H_i using sample weights λ_i. Given this formalism, we can build on existing tools for weighted classification to learn $\hat{\pi}$; see Zhou,

[18] Recall that $\tau_m(\cdot)$ does not depend on w in the case of binary treatments, and we omit the redundant argument in this case.

[19] In a recent advance, Farrell, Liang, and Misra (2020) established conditions under which deep neural networks can be shown to provably satisfy the conditions required by Assumption 2. Thus, depending on the statistical setting and the chosen architecture, deep neural networks could either be seen as a formally validated alternative to sieve-type methods or as heuristic method.

Athey, and Wager (2018) for a further discussion.[20] In all our experiments, we set Π to be a class of finite-depth decision trees (see Section 2.2 for a definition), and solve the optimization problem in Step 3 using our companion `R`-package `policytree` (Sverdrup, Kanodia, Zhou, Athey, and Wager (2020), R Core Team (2019)); see Zhou, Athey, and Wager (2018) for further details and motivation behind the computational strategy taken in this package.

5.1. *The California GAIN Program*

The Greater Avenues for Independence (GAIN) program, started in 1986, is a welfare-to-work program that provides participants with a mix of educational resources and job search assistance. Between 1988 and 1993, the Manpower Development Research Corporation conducted a randomized study to evaluate the program. As described in Hotz, Imbens, and Klerman (2006), randomly chosen registrants were eligible to receive GAIN benefits immediately, whereas others were embargoed from the program until 1993. All experimental subjects were followed for a 9-year post-randomization period and, as documented by Hotz, Imbens, and Klerman (2006), eligibility for GAIN had a significant impact on mean quarterly income averaged over this 9-year period.

Our current question is whether we can find ways to prioritize treatment to some subgroups of GAIN registrants particularly likely to benefit from it. We consider data from four counties, Alameda, Riverside, Los Angeles, and San Diego, resulting in $n = 19,170$ observations, and use $p = 28$ covariates, including demographics, education, and per-quarter earnings for 10 quarters preceding treatment. As in Hotz, Imbens, and Klerman (2006), we use average quarterly income over the 9-year post-randomization period (in \$1000s) as our outcome.

Each county participating in the GAIN evaluation conducted its own randomized controlled trial, and the counties had considerable freedom in how they carried out the randomization. In particular, counties had flexibility in choosing whom to enroll in the randomized trial, and which fraction of participants to randomize into treatment. The data reflects this heterogeneity in study specifications: The per-county average outcome for controls varied from 0.64 to 1.04 thousand dollars per quarter, while the per-county fraction of treated units varied from 0.50 to 0.86.

We use this dataset to design a semisynthetic observational study by pooling the data from all four counties under consideration. Because the mean control outcome and treatment fraction vary from county to county (and are in fact correlated), we expect that an uncorrected analysis of the pooled data would suffer from confounding. In an attempt to correct for the confounding that arises from pooling, we pursue a selection-on-observables strategy, and assume that controlling for the $p = 28$ covariates described above is enough to correct for the different study specifications used in different counties.

Our method starts by computing doubly robust scores for the treatment effect, and learning policies by empirical maximization as in (2). We use the augmented inverse-propensity weighted scores of Robins, Rotnitzky, and Zhao (1994), with nuisance component estimates from generalized random forests (Athey, Tibshirani, and Wager (2019),

[20]Some popular approaches for solving problems of the form (38) include best-subset empirical risk minimization (Chen and Lee (2018)) and optimal trees (Bertsimas and Dunn (2017)). Due to the computational difficulty of solving the problem (38) exactly, it may also be of interest to consider the empirical performance of alternative methods that solve an approximation to our weighted classification problem, for example, support vector machines (Cortes and Vapnik (1995)) or recursive partitioning (Breiman et al. (1984)). However, we caution that our formal results only apply to methods that solve the problem (38) exactly; see Wager (2019) for further discussion.

Breiman (2001)),[21]

$$\hat{\pi} = \operatorname{argmax}\left\{ \frac{1}{n}\sum_{i=1}^{n}\big(2\pi(X_i)-1\big)\big(\widehat{\Gamma}_i - C\big) : \pi \in \Pi \right\}, \tag{39}$$

$$\widehat{\Gamma}_i = \hat{\tau}^{(-i)}(X_i) + \frac{W_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\big(1-\hat{e}^{(-i)}(X_i)\big)}\Big(Y_i - \hat{f}^{(-i)}(X_i) - \big(W_i - \hat{e}^{(-i)}(X_i)\big)\hat{\tau}^{(-i)}(X_i)\Big), \tag{40}$$

where $\hat{f}(x)$ and $\hat{e}(x)$ are random forest estimates of $\mathbb{E}[Y_i|X_i = x]$ and $\mathbb{E}[W_i|X_i = x]$, respectively, $\hat{\tau}(\cdot)$ is an causal forest[22] estimate of the conditional average treatment effect, and C is a parameter measuring the cost of treatment. Tuning parameters for all forests were selected by leave-one-out cross-validation.[23] Here, we set $C = 0.14$ to roughly match the average treatment effect with the goal of ensuring that the optimal treatment rule is not trivial (i.e., we can only achieve nonzero utility gains by exploiting treatment heterogeneity).

Before starting to optimize policies, we first run a brief sanity check on our selection-on-observables strategy, and confirm the ability of estimators that build on this assumption to accurately recover the average treatment effect we would get using a proper randomization-based estimator that does not pool data across counties. The natural doubly robust estimator of the average treatment effect in our setting is $\hat{\theta}_{\mathrm{DR}} = \sum_{i=1}^{n}\widehat{\Gamma}_i/n$, with scores $\widehat{\Gamma}_i$ as in (40). We compare it to a naive difference-in-means estimator $\hat{\theta}_{\mathrm{DM}} = \mathrm{avg}\{Y_i : W_i = 1\} - \mathrm{avg}\{Y_i : W_i = 0\}$ that does not attempt to correct for bias due to pooling, and to an "oracle" doubly robust estimator that does not estimate propensity scores from covariates but instead uses the true per-county treated fractions: $\hat{\theta}^*_{\mathrm{DR}} = \sum_{i=1}^{n}\widehat{\Gamma}^*_i/n$ with

$$\begin{aligned}\widehat{\Gamma}^*_i &= \hat{\tau}^{(-i)}(X_i) + \frac{W_i - \hat{e}^*_i}{\hat{e}^*_i\big(1-\hat{e}^*_i\big)}\Big(Y_i - \hat{f}^{(-i)}(X_i) - \big(W_i - \hat{e}^*_i\big)\hat{\tau}^{(-i)}(X_i)\Big),\\ \hat{e}^*_i &= \sum_{j=1}^{n} W_j 1\big(\{G_j = G_i\}\big) \Big/ \sum_{j=1}^{n} 1\big(\{G_j = G_i\}\big),\end{aligned} \tag{41}$$

[21]The one major deviation between how we compute scores below and the assumptions of Theorem 1 is that, here, we use leave-one-out (or out-of-bag) estimates for $\tau(X_i)$, etc., whereas Theorem 1 assumed K-fold estimation. The reason for this choice is that, as discussed in Breiman (2001), random forests are particularly well suited for leave-one-out estimation, and allow the analyst to obtain such estimates at essentially no additional computational cost.

[22]Random forests are a type of adaptive nearest neighbor estimator that use an ensemble of trees to define a relevant neighborhood function for each query point; see Athey, Tibshirani, and Wager (2019) for a discussion. Causal forests use the adaptive neighborhood function implied by a forest to fit a partially linear model using the method of Robinson (1988); see Nie and Wager (2020) for formal results motivating the use of local partially linear modeling for heterogeneous treatment effect estimation, and Section 1.3 of Athey and Wager (2019) for a discussion of how this partially linear modeling is carried out in causal forests. We emphasize that, for our purposes, random forests are simply used as a convenient nonparametric estimator of relevant nuisance components, specifically $f(x)$ and $e(x)$ here, and could seamlessly be replaced with other methods such as boosting or neural networks. The shape of the learned policy $\hat{\pi}$ is determined in the optimization step 3, which only depends on the random forests through the predictions used to form doubly robust scores $\widehat{\Gamma}_i$.

[23]The regression surfaces $\hat{f}(x)$ and $\hat{e}(x)$ were tuned to optimize mean-squared error. As advocated in Nie and Wager (2020), the conditional average treatment effect function was tuned to optimize the error of a local residual-on-residual regression.

TABLE I

OUTCOME IS MEAN QUARTERLY INCOME (IN $1000) AVERAGED OVER 9 YEARS POST-INTERVENTION. DIFFERENCES IN MEAN RESPONSES BETWEEN WHITE AND NONWHITE RESPONDENTS ARE BOTH SIGNIFICANT AT THE 0.05 LEVEL USING A WELCH TWO-SAMPLE t-TEST

	Nonwhite	White
Fraction treated	76%	81%
Mean control outcome	0.79	0.90

where $G_i \in \{\text{Alameda, Riverside, Los Angeles, San Diego}\}$ denotes the county-membership of the ith sample. Because $\hat{\theta}^*_{\text{DR}}$ uses the true per-county treatment fractions $\hat{e}^*_i$ and estimates nuisance components using cross-fitting, the point estimates will be $\sqrt{n}$-consistent and the associated confidence intervals asymptotically valid essentially without assumptions (Rothe (2018), Wager, Du, Taylor, and Tibshirani (2016)). The resulting point estimates for the average treatment effect (± 1 standard error) are: $\hat{\theta}_{\text{DR}} = 0.141 \pm 0.026$ for the feasible doubly-robust estimator, $\hat{\theta}^*_{\text{DR}} = 0.146 \pm 0.028$ for the oracle doubly-robust estimator, and $\hat{\theta}_{\text{DM}} = 0.208 \pm 0.028$ for the naive difference in means. Thus, it appears that pooling county information results in confounding, but that controlling for available covariates helps.

We now move to learning a policy $\hat{\pi}$. In doing so, however, we note that caution is warranted because we have measured features pertaining to race, ethnicity, age, and gender. On the one hand, there may be legal restrictions on the use of these features for treatment allocation but, on the other hand, they appear to act as counfounders. For example, as shown in Table I, white GAIN registrants were randomized to treatment at higher rates than nonwhite registrants, and also white controls had higher outcomes than nonwhite controls. Our approach allows us to seamlessly use such sensitive variables for deconfounding without using them for policy allocation: We use these variables when estimating the nuisance components in (40), but then omit them from the maximization step (39) that produces the policy.

For our policy class Π, we consider decision trees of depth either 1 or 2. The learned decision rules are shown in Figure 1. Interestingly, the depth-1 and 2 trees make the same decisions for the roughly 3/4 of GAIN registrants who were paid 3 quarters prior to randomization, but the depth-2 tree chooses to switch to a different rule for those who were not paid 3 quarters prior.

In order to choose tree depth and, more broadly, to evaluate the accuracy of the policy learning procedure, we recommend cross-validation. We randomly divide the data into K

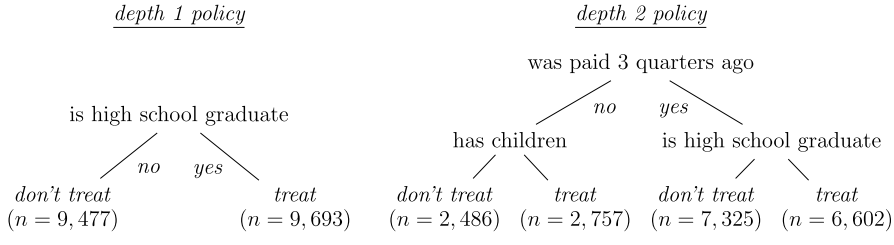

FIGURE 1.—Example of optimal depth-1 and -2 policy trees learned by optimizing the augmented inverse-propensity weighting loss function.

TABLE II

ESTIMATE OF THE UTILITY IMPROVEMENT OF VARIOUS POLICIES OVER A RANDOM ASSIGNMENT BASELINE, ±1 STANDARD ERROR. THE PLUG-IN POLICY SIMPLY THRESHOLDS CAUSAL FOREST PREDICTIONS AT $\hat{\tau}^{(-i)}(X_i) > C$, THE INVERSE-PROPENSITY WEIGHTED TREES (IPW) ARE FOLLOWING KITAGAWA AND TETENOV (2018), AND THE TREES SCORED VIA AUGMENTED INVERSE PROPENSITY-WEIGHTING (AIPW) ARE INSTANCES OF THE METHOD STUDIED HERE. THE LEFT COLUMN ESTIMATES IMPROVEMENT VIA (42), WHEREAS THE RIGHT COLUMN BRINGS IN COUNTY MEMBERSHIP INFORMATION TO OBTAIN A RANDOMIZATION-BASED ESTIMATOR OF IMPROVEMENT (43)

	Estimated Improvement	
Method	Fitted Propensities	True Propensities
Plug-in	0.077 ± 0.026	0.063 ± 0.028
IPW depth 1	0.065 ± 0.026	0.048 ± 0.028
IPW depth 2	0.043 ± 0.026	0.029 ± 0.028
AIPW depth 1	0.068 ± 0.026	0.050 ± 0.028
AIPW depth 2	0.091 ± 0.026	0.080 ± 0.028

folds $\mathcal{S}_k$, $k = 1, \ldots, K$ and, for each fold, learn a policy $\hat{\pi}^{(-k)}(\cdot)$ using all but the data in $\mathcal{S}_k$. Here, we use $K = 10$. Finally, we estimate improvement over a random baseline as

$$\widehat{A}_{CV} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} \left(2\hat{\pi}^{(-k)}(X_i) - 1\right) \widehat{\Gamma}_i. \tag{42}$$

Table II shows the estimated improvement of our depth-1 and -2 trees, as well as two baselines: A variant of the inverse-propensity weighted method of Kitagawa and Tetenov (2018) using the propensities used to construct (40), as well as a plug-in policy that does not obey our functional form restriction, and simply treats all samples with $\hat{\tau}^{(-i)}(X_i) > C$. Our depth-2 trees achieve markedly better performance than the depth-1 trees. Interestingly, the depth-2 tree is also competitive with the unconstrained plug-in estimator. Based on this analysis, we prefer the depth-2 tree in Figure 1.

One potential concern with this analysis is that our evaluation hinges on validity of the selection-on-observables assumption, as well as accuracy of the doubly robust scores $\widehat{\Gamma}_i$ from (40). To assuage this concern, we also computed a version of the improvement measure (42), but with scores $\widehat{\Gamma}_i^*$ computed using the true per-county treatment fractions as in (41):

$$\widehat{A}^*_{CV} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} \left(2\hat{\pi}^{(-k)}(X_i) - 1\right) \widehat{\Gamma}^*_i. \tag{43}$$

As seen in the second rightmost column of Table II, our feasible evaluation discussed above gave the correct ordering for the methods, but was somewhat optimistic in terms of the quality of the learned policies. The formal properties of treatment rules whose complexity is tuned via cross-validation are considered by Mbakop and Tabord-Meehan (2016).[24]

[24] Recall that cross-validation is a means of evaluating the quality of the policy learning procedure, not the decision that was produced by a specific realization of the procedure. If we want an accuracy assessment that is valid conditionally on the learned rule $\hat{\pi}(\cdot)$, one can either use a single test-train split, or use the more sophisticated data carving approach of Fithian, Sun, and Taylor (2014).

5.2. *Simulation Study With Binary, Endogenous Treatments*

In order to develop a richer quantitative understanding of the behavior of our method, we now turn to a simulation study. Here, we consider a setting with a binary, endogenous treatment W_i and a binary instrument Z_i and assume homogeneity as in (16). In this case, our method chooses the policy $\hat{\pi} = \text{argmax}\{\frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\widehat{\Gamma}_i : \pi \in \Pi\}$, where $\widehat{\Gamma}_i$ is a cross-fit doubly robust score with estimates of the compliance weights as in (17):

$$\begin{aligned}\widehat{\Gamma}_i &= \hat{\tau}^{(-i)}(X_i) \\ &\quad + \hat{g}^{(-i)}(X_i, Z_i)\big(Y_i - \hat{f}^{(-i)}(X_i) - \big(W_i - \hat{e}^{(-i)}(X_i)\big)\hat{\tau}^{(-i)}(X_i)\big), \\ \hat{g}^{(-i)}(X_i, Z_i) &= \frac{1}{\widehat{\Delta}^{(-i)}(X_i)}\frac{Z_i - \hat{z}^{(-i)}(X_i)}{\hat{z}^{(-i)}(X_i)\big(1-\hat{z}^{(-i)}(X_i)\big)},\end{aligned} \tag{44}$$

where $\Delta(x) = \mathbb{P}[W_i = 1|Z_i = 1, X_i = x] - \mathbb{P}[W_i = 1|Z_i = 0, X_i = x]$ is the conditional average effect of the instrument on the treatment, $z(x) = \mathbb{P}[Z_i = 1|X_i = x]$, $f(x) = \mathbb{E}[Y_i|X_i = x]$, $e(x) = \mathbb{P}[W_i = 1|X_i = x]$, and $\tau(x)$ is the conditional average treatment effect as specified in (16). We estimate all nuisance components via random forest methods with the package `grf`, and use an instrumental forest for $\tau(\cdot)$, a causal forest for $\Delta(\cdot)$, and a regression forest for $f(\cdot)$, $e(\cdot)$ and $z(\cdot)$.

In this simulation experiment, we generate data independently as follows, for various choices of n and $\tau(\cdot)$:

$$\begin{aligned} X &\sim \mathcal{N}(0, \mathcal{I}_{10\times 10}), \qquad Z|X \sim \text{Bernoulli}\big(1/\big(1+e^{-X_3}\big)\big), \\ \varepsilon|X, Z &\sim \mathcal{N}(0,1), \qquad Q|X, Z, \varepsilon \sim \text{Bernoulli}\big(1/\big(1+e^{-\varepsilon - X_4}\big)\big), \\ W &= Q \wedge Z, \qquad Y = (X_3 + X_4)_+ + W\tau(X) + \varepsilon. \end{aligned} \tag{45}$$

Note that W is in fact endogenous, because Q (and thus also W) is more likely to be 1 when the noise term ε is large. Given this setup, we consider $\tau(\cdot)$ functions

$$\tau(x) = \big((x_1)_+ + (x_2)_+ - 1\big)/2 \quad \text{and} \tag{46}$$

$$\tau(x) = \text{sign}(x_1 x_2)/2. \tag{47}$$

In both cases, we learn $\pi(\cdot)$ over the class Π of depth-2 trees and note that best nonparametric policy $\pi^*(x) = 1(\{\tau(x) > 0\})$ belongs to Π in case (47) but not in case (46).

In Figure 2, we display the improvement $A(\pi) = \mathbb{E}[(2\pi(X_i)-1)\tau(X_i)]$ of our learned policies relative to a random assignment baseline, for different values of n. Overall, we see that the regret of the learned policies improves with n, and approaches best-in-class regret as n gets large. We also note an interesting difference in the behavior of the learned rules in settings (46) and (47). In the first case, $\tau(\cdot)$ is continuous, and regret improves smoothly with sample size. Conversely, in the second case where $\tau(\cdot)$ has sharp jumps, we observe something of a phase transition between $n = 2000$ and $n = 4000$, as our trees become able to consistently make splits that roughly match the jumps in $\tau(\cdot)$.

6. DISCUSSION

In this paper, we proposed an approach to policy learning in the observational study setting that builds on classical ideas for semiparametrically efficient treatment effect estimation. Our main result is that doubly robust estimators of average treatment effects

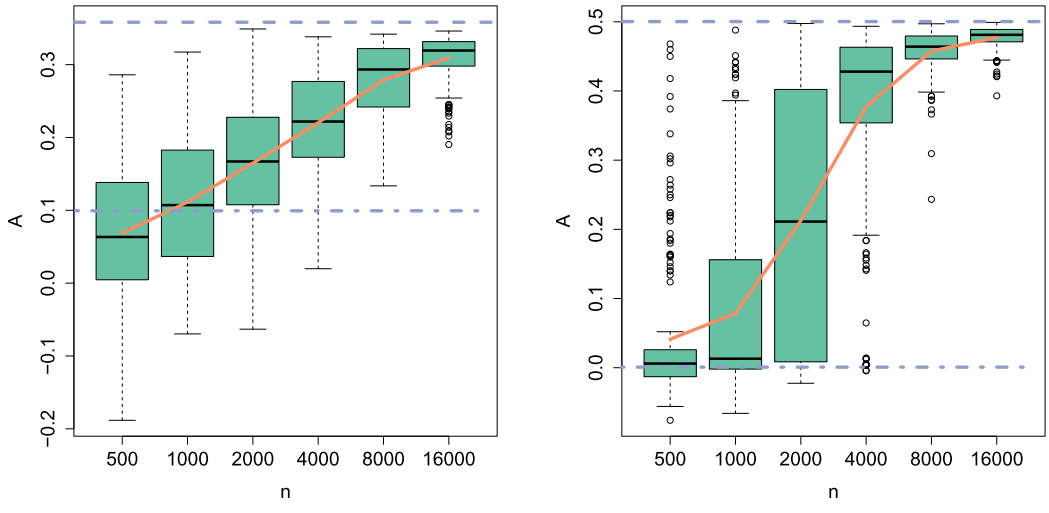

FIGURE 2.—Distribution of the improvement $A(\pi) = \mathbb{E}[(2\pi(X_i) - 1)\tau(X_i)]$ for policies learned by optimizing the scores (44) over the class Π of depth-2 trees, for different values of sample size n. Each box plot summarizes the distribution of $A(\hat{\pi})$ over 200 simulation replications, while the solid line shows the average of $A(\hat{\pi})$. The lower horizontal line shows $A(\pi)$ for the best policy that does not use the features X_i (i.e., either always treat or never treat), and the upper horizontal line shows the supremum of $A(\pi)$ over the class Π.

can be adapted for policy evaluation, and that the policy that maximizes the resulting doubly robust value estimate over a prespecified class Π satisfies rate-optimal guarantees for minimax regret. Our approach decouples estimation of nuisance components used for the doubly robust scores from optimization of the doubly robust value function, and thus allows practitioners flexibility in how they implement each step.

Our formal discussion focused on regret bounds for policy learning. A natural follow-up question is to ask for confidence sets guaranteed to contain an optimal policy: For example, if Π is the set of depth-L decision trees, can we identify a subset of Π guaranteed to contain a value-maximizing policy in Π with high probability? Some early results in this direction are reported by Rai (2018). Meanwhile, Armstrong and Shen (2013) considered the related task of identifying a subset of the population we are confident will benefit from the policy intervention.

Another natural direction to extend our results is toward dynamic decision making problems, where the policy maker needs to make a sequence of decisions, potentially depending on time-varying covariates. The problem of doubly robust policy evaluation in this setting has been considered by Thomas and Brunskill (2016) and Zhang, Tsiatis, Laber, and Davidian (2013). Nie, Brunskill, and Wager (2019) proposed a method for learning observational stopping rules from observational data that is both computationally feasible and robust to confounding. Obtaining a more comprehensive landscape of the problem of dynamic policy learning in observational studies would be of considerable interest.

Finally, all results presented here relied on point-identification of treatment effects, either via a selection on observables assumption or via an instrument that satisfies conditional homogeneity. Some applications, however, do not allow for such clean assumptions, and thus call for methods for policy learning that are robust to failures of identifying assumptions. Kallus and Zhou (2020) considered the problem of policy learning under an

approximate selection-on-observables assumption in the sense of Rosenbaum (2002). It would also be of interest to study what can be done if we only have access to a monotone instrument, as in Manski and Pepper (2000).

REFERENCES

ABADIE, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113 (2), 231–263. [142]

AI, C., AND X. CHEN (2007): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models With Different Conditioning Variables," *Journal of Econometrics*, 141 (1), 5–43. [141]

ARMSTRONG, T. B., AND M. KOLESÁR (2017): "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness," arXiv preprint arXiv:1712.04594. [137]

ARMSTRONG, T. B., AND S. SHEN (2013): "Inference on Optimal Treatment Assignments," Technical report, Cowles Foundation for Research in Economics, Yale University. [136,157]

ARONOW, P. M., AND A. CARNEGIE (2013): "Beyond Late: Estimation of the Average Treatment Effect With an Instrumental Variable," *Political Analysis*, 21 (4), 492–506. [142]

ATHEY, S., AND G. IMBENS (2016): "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113 (27), 7353–7360. [137]

ATHEY, S., AND S. WAGER (2019): "Estimating Treatment Effects With Causal Forests: An Application," *Observational Studies*, 5, 36–51. [153]

——— (2021): "Supplement to 'Policy Learning With Observational Data'," *Econometrica Supplemental Material*, 89, https://doi.org/10.3982/ECTA15732. [143,145]

ATHEY, S., J. TIBSHIRANI, AND S. WAGER (2019): "Generalized Random Forests," *The Annals of Statistics*, 47 (2), 1148–1178. [137,152,153]

BAN, G.-Y., AND C. RUDIN (2019): "The Big Data Newsvendor: Practical Insights From Machine Learning," *Operations Research*, 67 (1), 90–108. [137]

BARTLETT, P. L., AND S. MENDELSON (2002): "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results," *Journal of Machine Learning Research*, 3, 463–482. [145,146]

——— (2006): "Empirical Minimization," *Probability Theory and Related Fields*, 135 (3), 311–334. [150]

BARTLETT, P. L., O. BOUSQUET, AND S. MENDELSON (2005): "Local Rademacher Complexities," *Annals of Statistics*, 33 (4), 1497–1537. [146]

BERTSIMAS, D., AND J. DUNN (2017): "Optimal Classification Trees," *Machine Learning*, 106 (7), 1039–1082. [152]

BERTSIMAS, D., AND N. KALLUS (2020): "From Predictive to Prescriptive Analytics," *Management Science*, 66 (3), 1025–1044. [137]

BEYGELZIMER, A., AND J. LANGFORD (2009): "The Offset Tree for Learning With Partial Labels," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 129–138. [137,151]

BHATTACHARYA, D., AND P. DUPAS (2012): "Inferring Welfare Maximizing Treatment Assignment Under Budget Constraints," *Journal of Econometrics*, 167 (1), 168–196. [136]

BICKEL, P., C. KLAASSEN, Y. RITOV, AND J. WELLNER (1998): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag. [137,140]

BREIMAN, L. (2001): "Random Forests," *Machine Learning*, 45 (1), 5–32. [153]

BREIMAN, L., J. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE (1984): *Classification and Regression Trees*. CRC press. [143,152]

CAPONNETTO, A., AND E. DE VITO (2007): "Optimal Rates for the Regularized Least-Squares Algorithm," *Foundations of Computational Mathematics*, 7 (3), 331–368. [140,151]

CHAMBERLAIN, G. (2011): "Bayesian Aspects of Treatment Choice," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. V. Dijk. Oxford University Press. [136]

CHEN, L.-Y., AND S. LEE (2018): "Best Subset Binary Prediction," *Journal of Econometrics*, 206 (1), 39–56. [152]

CHEN, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," *Handbook of Econometrics*, 6, 5549–5632. [140,151]

CHEN, X., H. HONG, AND A. TAROZZI (2008): "Semiparametric Efficiency in GMM Models With Auxiliary Data," *The Annals of Statistics*, 36 (2), 808–843. [137,140]

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21 (1), C1–C68. [135,140,143,147]

CHERNOZHUKOV, V., J. C. ESCANCIANO, H. ICHIMURA, W. K. NEWEY, AND J. M. ROBINS (2016): "Locally Robust Semiparametric Estimation," arXiv preprint arXiv:1608.00033. [134,135,137,139-141,143,147]
CORTES, C., AND V. VAPNIK (1995): "Support-Vector Networks," *Machine learning*, 20 (3), 273–297. [152]
CORTES, C., Y. MANSOUR, AND M. MOHRI (2010): "Learning Bounds for Importance Weighting," in *Advances in Neural Information Processing Systems*, 442–450. [146]
DEHEJIA, R. H. (2005): "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125 (1), 141–173. [136]
DEVROYE, L., AND G. LUGOSI (1995): "Lower Bounds in Pattern Recognition and Learning," *Pattern recognition*, 28 (7), 1011–1018. [150]
DUDÍK, M., J. LANGFORD, AND L. LI (2011): "Doubly Robust Policy Evaluation and Learning," in *Proceedings of the 28th International Conference on Machine Learning*, 1097–1104. [136,137]
FARRELL, M. H. (2015): "Robust Inference on Average Treatment Effects With Possibly More Covariates Than Observations," *Journal of Econometrics*, 189 (1), 1–23. [141]
FARRELL, M. H., T. LIANG, AND S. MISRA (2020): "Deep Neural Networks for Estimation and Inference," Econometrica. (forthcoming). [151]
FITHIAN, W., D. SUN, AND J. TAYLOR (2014): "Optimal Inference After Model Selection," arXiv preprint arXiv:1410.2597. [155]
GINÉ, E., AND V. KOLTCHINSKII (2006): "Concentration Inequalities and Asymptotic Results for Ratio Type Empirical Processes," *The Annals of Probability*, 34 (3), 1143–1216. [146]
HAHN, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66 (2), 315–331. [137,140]
HIRANO, K., AND J. R. PORTER (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77 (5), 1683–1701. [134,136,138,150]
HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71 (4), 1161–1189. [135,148]
HIRSHBERG, D. A., AND S. WAGER (2018): "Augmented Minimax Linear Estimation," arXiv preprint arXiv:1712.00038. [144]
HOTZ, V. J., G. W. IMBENS, AND J. A. KLERMAN (2006): "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program," *Journal of Labor Economics*, 24 (3), 521–566. [152]
HUDGENS, M. G., AND M. E. HALLORAN (2008): "Toward Causal Inference With Interference," *Journal of the American Statistical Association*, 103 (482), 832–842. [135]
IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62 (2), 467–475. [142]
IMBENS, G. W., AND D. B. RUBIN (2015): *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press. [135,144]
KALLUS, N. (2018): "Balanced Policy Evaluation and Learning," in *Advances in Neural Information Processing Systems*, 8895–8906. [136,137]
KALLUS, N., AND A. ZHOU (2020): "Confounding-Robust Policy Improvement," Management Science. (forthcoming). [157]
KASY, M. (2016): "Partial Identification, Distributional Preferences, and the Welfare Ranking of Policies," *Review of Economics and Statistics*, 98 (1), 111–131. [136]
KITAGAWA, T., AND A. TETENOV (2018): "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86 (2), 591–616. [134,136-138,141,146,150,151,155]
LUEDTKE, A., AND A. CHAMBAZ (2020): "Performance Guarantees for Policy Learning," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56 (3), 2162–2188. [137,150]
LUEDTKE, A. R., AND M. J. VAN DER LAAN (2016): "Statistical Inference for the Mean Outcome Under a Possibly Non-Unique Optimal Treatment Strategy," *The Annals of Statistics*, 44 (2), 713–742. [137]
MANSKI, C. F. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72 (4), 1221–1246. [134,136,138]
——— (2009): *Identification for Prediction and Decision*. Harvard University Press. [134-136,138]
——— (2013): "Identification of Treatment Response With Social Interactions," *The Econometrics Journal*, 16 (1), S1–S23. [135]
MANSKI, C. F., AND J. V. PEPPER (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68 (4), 997–1010. [158]
MAURER, A., AND M. PONTIL (2009): "Empirical Bernstein Bounds and Sample Variance Penalization," in *Conference on Learning Theory*. [146]
MBAKOP, E., AND M. TABORD-MEEHAN (2016): "Model Selection for Treatment Choice: Penalized Welfare Maximization," arXiv preprint arXiv:1609.03167. [143,155]

NEWEY, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62 (6), 1349–1382. [137,140,144]
NEYMAN, J. (1923): "Sur les Applications de la Théorie des Probabilités aux Experiences Agricoles: Essai des Principes," *Roczniki Nauk Rolniczych*, 10, 1–51. [138]
NICKL, R., AND B. M. PÖTSCHER (2007): "Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov-and Sobolev-Type," *Journal of Theoretical Probability*, 20 (2), 177–199. [149]
NIE, X., AND S. WAGER (2020): "Quasi-Oracle Estimation of Heterogeneous Treatment Effects," Biometrika. (forthcoming). [137,153]
NIE, X., E. BRUNSKILL, AND S. WAGER (2019): "Learning When-to-Treat Policies," Journal of the American Statistical Association. (forthcoming). [157]
POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57 (6), 1403–1430. [138,141]
QIAN, M., AND S. A. MURPHY (2011): "Performance Guarantees for Individualized Treatment Rules," *Annals of Statistics*, 39 (2), 1180–1210. [137]
R CORE TEAM (2019): *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/. [152]
RAI, Y. (2018): "Statistical Inference for Treatment Assignment Policies," Unpublished Manuscript. [157]
ROBINS, J., AND A. ROTNITZKY (1995): "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90 (1), 122–129. [137,140]
ROBINS, J., A. ROTNITZKY, AND L. P. ZHAO (1994): "Estimation of Regression Coefficients When Some Regressors Are not Always Observed," *Journal of the American Statistical Association*, 89 (427), 846–866. [136,141,152]
ROBINS, J. M., L. LI, R. MUKHERJEE, E. T. TCHETGEN, AND A. VAN DER VAART (2017): "Minimax Estimation of a Functional on a Structured High-Dimensional Model," *The Annals of Statistics*, 45 (5), 1951–1987. [144]
ROBINSON, P. M. (1988): "Root-n-Consistent Semiparametric Regression," *Econometrica*, 56 (4), 931–954. [153]
ROSENBAUM, P. R. (2002): *Observational Studies*. Springer Science & Business Media. [158]
ROTHE, C. (2018): "Flexible Covariate Adjustments in Randomized Experiments", https://madoc.bib.uni-mannheim.de/52249/. [154]
RUBIN, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66 (5), 688–701. [138]
SCHICK, A. (1986): "On Asymptotically Efficient Estimation in Semiparametric Models," *The Annals of Statistics*, 14 (3), 1139–1151. [140,144]
STOYE, J. (2009): "Minimax Regret Treatment Choice With Finite Samples," *Journal of Econometrics*, 151 (1), 70–81. [134,136,138]
——— (2012): "Minimax Regret Treatment Choice With Covariates or With Limited Validity of Experiments," *Journal of Econometrics*, 166 (1), 138–156. [134,136]
SVERDRUP, E., A. KANODIA, Z. ZHOU, S. ATHEY, AND S. WAGER (2020): "Policytree: Policy Learning via Doubly Robust Empirical Welfare Maximization Over Trees," *Journal of Open Source Software*, 5 (50), 2232. [152]
SWAMINATHAN, A., AND T. JOACHIMS (2015): "Batch Learning From Logged Bandit Feedback Through Counterfactual Risk Minimization," *Journal of Machine Learning Research*, 16, 1731–1755. [134,137,146]
TETENOV, A. (2012): "Statistical Treatment Choice Based on Asymmetric Minimax Regret Criteria," *Journal of Econometrics*, 166 (1), 157–165. [136]
THOMAS, P., AND E. BRUNSKILL (2016): "Data-Efficient off-Policy Policy Evaluation for Reinforcement Learning," in *International Conference on Machine Learning*, 2139–2148. [157]
VAN DER LAAN, M. J., AND S. ROSE (2011): *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media. [135,140]
VAN DER LAAN, M. J., S. DUDOIT, AND A. W. VAN DER VAART (2006): "The Cross-Validated Adaptive Epsilon-Net Estimator," *Statistics & Decisions*, 24 (3), 373–395. [136]
VAN DER LAAN, M. J., E. C. POLLEY, AND A. E. HUBBARD (2007): "Super Learner," *Statistical applications in genetics and molecular biology*, 6 (1), 25. [151]
VAPNIK, V. (2000): *The Nature of Statistical Learning Theory*. Springer Information Science and Statistics. [142]
WAGER, S. (2019): "On Regression Tables for Policy Learning: Comment on a Paper by Jiang, Song, li and Zeng," *Statistica Sinica*, 29 (4), 1678–1684. [152]
WAGER, S., W. DU, J. TAYLOR, AND R. J. TIBSHIRANI (2016): "High-Dimensional Regression Adjustments in Randomized Experiments," *Proceedings of the National Academy of Sciences*, 113 (45), 12673–12678. [154]

WAINWRIGHT, M. J. (2019): *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Vol. 48. Cambridge University Press. [142]

ZHANG, B., A. A. TSIATIS, M. DAVIDIAN, M. ZHANG, AND E. LABER (2012): "Estimating Optimal Treatment Regimes From a Classification Perspective," *Stat*, 1 (1), 103–114. [136,137,151]

ZHANG, B., A. A. TSIATIS, E. B. LABER, AND M. DAVIDIAN (2013): "Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions," *Biometrika*, 100 (3), 681–694. [157]

ZHAO, Y., D. ZENG, A. J. RUSH, AND M. R. KOSOROK (2012): "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107 (499), 1106–1118. [134,137,151]

ZHENG, W., AND M. J. VAN DER LAAN (2011): "Cross-Validated Targeted Minimum-Loss-Based Estimation," in *Targeted Learning*. Springer, 459–474. [144]

ZHOU, Z., S. ATHEY, AND S. WAGER (2018): "Offline Multi-Action Policy Learning: Generalization and Optimization," arXiv preprint arXiv:1810.04778. [143,152]

Co-editor Ulrich K. Müller handled this manuscript.

Manuscript received 2 October, 2017; final version accepted 3 September, 2020; available online 4 September, 2020.