Efficient Policy Learning

Susan Athey
athey@stanford.edu

Stefan Wager
swager@stanford.edu

Draft version October 2017

Abstract

We consider the problem of using observational data to learn treatment assignment policies that satisfy certain constraints specified by a practitioner, such as budget, fairness, or functional form constraints. This problem has previously been studied in economics, statistics, and computer science, and several regret-consistent methods have been proposed. However, several key analytical components are missing, including a characterization of optimal methods for policy learning, and sharp bounds for minimax regret. In this paper, we derive lower bounds for the minimax regret of policy learning under constraints, and propose a method that attains this bound asymptotically up to a constant factor. Whenever the class of policies under consideration has a bounded Vapnik-Chervonenkis dimension, we show that the problem of minimax-regret policy learning can be asymptotically reduced to first efficiently evaluating how much each candidate policy improves over a randomized baseline, and then maximizing this value estimate. Our analysis relies on uniform generalizations of classical semiparametric efficiency results for average treatment effect estimation, paired with sharp concentration bounds for weighted empirical risk minimization that may be of independent interest.

Keywords: asymptotic theory, double machine learning, double robustness, empirical welfare maximization, empirical process, 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. There is an increasing number of application areas with rich, observational datasets that can be used to learn personalized treatment rules. Moreover, technology companies and educational institutions have begun to introduce explicit randomization into their systems in order to enable policy learning using routinely logged data. The goal of this paper is to develop an understanding of how such observational datasets can be used to learn policies in a way that uses the available data as efficiently as possible.

We are grateful for helpful conversations with colleagues including Guido Imbens, Michael Kosorok, Alexander Luedtke, Eric Mbakop, Alexander Rakhlin, James Robins, Max Tabord-Meehan and Zhengyuan Zhou, and for feedback from seminar participants at a variety of universities and workshops.

The minimax regret criterion for policy learning As in Manski (2004, 2009), we formalize the task of optimal policy learning via a utilitarian minimax regret criterion. Following the potential outcomes model (Neyman, 1923; Rubin, 1974), we posit the existence of a data-generating distribution on triples $\{X_i,\, Y_i(-1), Y(+1)\} \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}$, where the X_i are observable characteristics of the i-th unit, and the $Y_i(w)$, $w \in \{\pm 1\}$, denote the utility the i-th unit would have experienced given treatment w. A treatment assignment policy $\pi(\cdot)$ is a mapping from a unit's characteristics X_i to a specification $\pi(X_i) \in \{\pm 1\}$ of which treatment the unit should receive. For a class of candidate policies Π, we define an optimal policy π^* as one that maximizes expected utility among all policies $\pi \in \Pi$, and regret as the difference between the expected utility of a given policy π and that of the optimal policy,[1]

$$\pi^* \in \text{argmax}_{\pi \in \Pi} \left\{ \mathbb{E}\left[Y_i\left(\pi(X_i)\right) \right] \right\}, \quad R(\pi) = \mathbb{E}\left[Y_i\left(\pi^*(X_i)\right) \right] - \mathbb{E}\left[Y_i\left(\pi(X_i)\right) \right]. \tag{1}$$

Given this setup, we assume that a practitioner seeks to learn a policy $\hat{\pi}$ with the best possible upper bound on regret $R(\hat{\pi})$. This minimax regret criterion underlies several key developments in statistical decision theory, including the work of Lai and Robbins (1985) on multi-armed bandits. It is usually attributed to Savage (1951), who proposed it in a review of Wald (1950).

Throughout this paper, we study a setup where the practitioner chooses $\hat{\pi}$ after observing $i = 1, 2, ..., n$ independent and identically distributed units of the form $\{X_i,\, Y_i,\, W_i\} \in \mathcal{X} \times \mathbb{R} \times \{\pm 1\}$, where $\{X_i,\, Y_i(-1), Y(+1)\}$ are drawn from the population specified in the above paragraph, and $Y_i = Y_i(W_i)$ for a treatment assignment variable W_i. We take W_i to be an unconfounded random variable with overlap (Rosenbaum and Rubin, 1983),

$$\{Y_i(-1),\, Y_i(+1)\} \perp\!\!\!\perp W_i \,\big|\, X_i = x, \qquad \left| \mathbb{E}\left[W_i \,\big|\, X_i = x \right] \right| \leq 1 - 2\eta \tag{2}$$

for some $\eta > 0$ and all $x \in \mathcal{X}$. These assumptions enable identification of causal effects in observational studies (Imbens and Rubin, 2015). Our goal is to use such observational data to learn policies $\hat{\pi}$ that have low regret $R(\hat{\pi})$ with high probability, while only making generic regularity assumptions about the joint distribution of $\{X_i,\, Y_i(-1),\, Y_i(+1),\, W_i\}$.[2]

Now, despite the seeming simplicity of the problem outlined above, the criterion (1) can motivate substantively different statistical tasks depending on how we specify the the class of allowable policies Π.[3] At one end of the spectrum, some authors place no restrictions on this class of policies, and let Π consist of all (sufficiently regular) functions from $\mathcal{X}$ to $\{\pm 1\}$. Manski (2004) considers this setting in the case where $\mathcal{X}$ has finite support, and

[1] All results presented in this paper will also hold in the case where no optimal policy exists, provided we make appropriate notational adjustments; for example, we would need to redefine regret as $R(\pi) = \sup\{\mathbb{E}[Y_i(\bar{\pi}\,(X_i))] : \bar{\pi} \in \Pi\} - \mathbb{E}\left[Y_i\left(\pi\,(X_i)\right) \right]$. However, to simplify our exposition, we state our results in the case where π^* exists.

[2] Our setup, where we want to learn a policy based on already collected observational data, is in contrast to the online "contextual bandit" setup where the practitioner seeks to learn a decision rule while actively making treatment allocation decisions for incoming subjects (e.g., Agarwal et al., 2014; Auer et al., 2002; Bastani and Bayati, 2015; Lai and Robbins, 1985; Perchet and Rigollet, 2013; Rakhlin and Sridharan, 2016). The contextual bandit problem is quite different from ours: On one hand, it is harder because of an exploration/exploitation trade-off that arises in sequential trials; on the other hand, it is easier, because treatment propensities are known (since they were explicitly specified during the sequential trial). In the machine learning literature, our setup is sometimes called the "offline bandit" problem.

[3] The results in this paper are all built using a frequentist minimax problem setting. For a discussion of Bayesian policy learning, see, e.g., Chamberlain (2011) or Dehejia (2005). For papers that consider alternative welfare criteria, see Kitagawa et al. (2017), who study an equality-weighted version of the welfare criterion; Tetenov (2012), who considers asymmetric criterion across type I and type II errors; and also Kasy (2016) examines measure for comparing distributions of policy outcomes.

shows that the simple decision rule obtained by thresholding an efficient estimator of the conditional average treatment effect $\tau(x) = \mathbb{E}\left[Y(+1) - Y(-1) \,\middle|\, X = x\right]$ is asymptotically minimax optimal. His result is further refined by Hirano and Porter (2009), who show that such thresholding rules are also optimal when $\mathcal{X}$ is a continuum under local asymptotics as motivated by Le Cam (1986),[4] and by Stoye (2009, 2012) who derives exact minimax treatment allocation rules in the setting of Manski (2004).

At the other end of the spectrum, one may also consider the case where Π only consists of two possible policies: treat everyone ($\pi(x) = +1$ for all $x \in \mathcal{X}$) or treat no one ($\pi(x) = -1$ for all $x \in \mathcal{X}$). In this case, the problem of policy learning becomes effectively equivalent to the problem of estimating an average treatment effect under unconfoundedness,[5] which is the topic of another well developed literature (see Imbens and Rubin (2015) for a review), with notable contributions from Hahn (1998), Hirano et al. (2003), Robins and Rotnitzky (1995) and Robins et al. (2017), and recent extensions to high-dimensional problems by Athey et al. (2016a), Belloni et al. (2017) and Farrell (2015).

In this paper, we are most interested in the intermediate case where Π is neither unconstrained nor binary: For example, we might ask for $\pi(x)$ to be a sparse linear function of x, or a fixed-depth decision tree, possibly incorporating constraints such as budget constraints on the fraction of subjects receiving the treatment. As we discuss in more detail in Section 1.1, there are a variety of motivations for limiting the form of the assignment rule, from regulatory restrictions or costs for including certain covariates in assignment, to explainability or to simplicity of implementation. This intermediate "structured" setting appears to be considerably richer than the two extreme ones discussed above, and we are still far from having an exact optimality theory for it—despite considerable work in economics (Kitagawa and Tetenov, 2015; Mbakop and Tabord-Meehan, 2016), statistics (Luedtke and Chambaz, 2017; Qian and Murphy, 2011; Zhang et al., 2012; Zhao et al., 2012; Zhou et al., 2017) and machine learning (Beygelzimer and Langford, 2009; Dudík et al., 2011; Swaminathan and Joachims, 2015). The objective of our paper is to tie the problem of policy learning over structured classes Π to that of semiparametrically efficient policy evaluation, thus providing a sharp characterization (up to constants) of the difficulty of this problem.[6]

Policy learning via empirical maximization In order to present our contribution, it is helpful to first review existing results. Given our goal of minimizing regret, a natural approach to policy learning is to optimize empirical regret estimates (e.g., Dudík et al., 2011; Kitagawa and Tetenov, 2015; Swaminathan and Joachims, 2015; Zhao et al., 2012).

[4] See also Hirano and Porter (2016), who consider extensions to panel data.

[5] Heuristically, the connection is obvious, because we want to treat everyone if and only if the average treatment effect is positive. To obtain a formal equivalence statement for the two statistical tasks, one could again rely on Le Cam-style local asymptotics as used by Hirano and Porter (2009).

[6] Following standard practice in this literature, we focus on establishing upper bounds rather than exact expressions for $R(\pi)$ (e.g., Kitagawa and Tetenov, 2015; Manski, 2004; Swaminathan and Joachims, 2015), for the simple reason that exact asymptotics for discrete optimization problems of this type are often intractable (see Bartlett and Mendelson (2006) for a discussion). There are of course some notable exceptions: Stoye (2009) uses a game-theoretic approach to derive exact minimax treatment allocation rules in the special case where $\mathcal{X}$ is discrete and Π is unrestricted (see also Stoye (2012)), while Hirano and Porter (2009) obtain exact asymptotics for regret using carefully set up Le Cam-style asymptotics, again provided that Π is unrestricted. The difference between our setting and the work of Hirano and Porter (2009) and Stoye (2009, 2012) is that we allow Π to have arbitrary structure, which makes exact analysis combinatorially intractable.

Procedurally, this amounts to specifying $\hat{\pi}$ as

$$\hat{\pi} = \text{argmax}_{\pi \in \Pi} \left\{ \widehat{Q}(\pi) \right\}, \text{ where } \frac{Q(\pi)}{2} = \mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right] - \frac{1}{2}\mathbb{E}\left[Y_i(-1) + Y_i(+1)\right] \tag{3}$$

and $\widehat{Q}(\pi)$ is some uniformly consistent estimator for $Q(\pi)$ over the class $\pi \in \Pi$. The quantity $Q(\pi)$ is interpreted as a "policy value" that measures the improvement of π over a randomized baseline. As a concrete example of this line of work, Kitagawa and Tetenov (2015) consider (3) with a $\widehat{Q}$-estimator obtained via inverse-propensity weighting and show that, if (2) holds, the utilities Y_i are uniformly bounded, $|Y_i| \leq M$, and the policy class Π is a Vapnik-Chervonenkis class with dimension $\text{VC}(\Pi)$ (Vapnik and Chervonenkis, 1971), then $\hat{\pi}$ satisfies the regret bound

$$R\left(\hat{\pi}\right) = \mathcal{O}_P\left(\frac{M}{\eta}\sqrt{\frac{\text{VC}\left(\Pi\right)}{n}}\right). \tag{4}$$

Now, although such results have the virtue of providing regret-consistent treatment allocation rules $\hat{\pi}$ and helpfully highlight the relationship between the difficulty of policy learning and the ratio $\text{VC}\left(\Pi\right)/n$, they are still far from providing an optimality theory for policy learning. The main problem is that the dependence on the data-generating distribution via the ratio M/η in the bound (4) is too loose to meaningfully guide specific modeling choices. Although most of the methods discussed above can be cast in the form (3), many of them use different choices[7] for $\widehat{Q}(\pi)$; and yet, rather disappointingly, (4) is the best bound available for any of these proposals.

Efficient policy learning The goal of this paper is to provide more clarity on how to build good estimators of the form (3). We introduce some key concepts by first analyzing the problem of evaluating a single policy π as accurately as possible. Using notation from (3), we see that

$$Q\left(\pi\right) = \mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right] - \mathbb{E}\left[Y_i\left(-\pi(X_i)\right)\right], \tag{5}$$

i.e., $Q\left(\pi\right)$ is the average treatment effect in a randomized controlled trial where the "treated" sample is assigned policy $\pi(\cdot)$, and the "control" sample is assigned the opposite policy $-\pi(\cdot)$. Given unconfoundedness (2), there is a large literature that develops a semiparametric efficient estimation theory for statistics like $Q(\pi)$ (Bickel et al., 1998; Hahn, 1998; Hirano et al., 2003; Newey, 1994; Robins and Rotnitzky, 1995; Robins et al., 1995), for any fixed policy π.

Our main result is that we can translate results about efficient policy evaluation into results about policy learning. Specifically, let $V(\pi)$ denote the semiparametrically efficient variance for estimating $Q(\pi)$. Furthermore, let $V_* := V(\pi^*)$ denote the semiparametrically efficient variance for evaluating the optimal policy π^*, and let V_{max} (formally defined in (14) below) denote a sharp bound on the worst case efficient variance $\sup_\pi V(\pi)$ for any policy π. Then, under regularity conditions, we propose a learning rule that yields a policy $\hat{\pi}$ with

[7] For example, Kitagawa and Tetenov (2015) and Swaminathan and Joachims (2015) rely on inverse-propensity weighting, but Dudík et al. (2011), Zhang et al. (2012), Zhou et al. (2017) and Zhao et al. (2012) all make different recommendations.

regret bounded by[8]

$$R\left(\hat{\pi}\right) = \mathcal{O}_P\left(\sqrt{V_* \left(1 + \log\left(\frac{V_{\max}}{V_*}\right)\right) \frac{\mathrm{VC}\left(\Pi\right)}{n}}\right). \tag{6}$$

We also develop regret bounds for non-parametric policy classes Π with a bounded entropy integral, such as finite-depth decision trees. Key components of our analysis include uniform concentration results for efficient policy evaluation, as well as sharp generalization bounds for weighted empirical risk minimization that may be of independent interest.[9]

This result has several implications. First, on a conceptual level, we note that when n is large, our bound (6) is strictly better than any other regret bound for policy learning previously proposed in the literature (Beygelzimer and Langford, 2009; Kitagawa and Tetenov, 2015; Swaminathan and Joachims, 2015; Zhao et al., 2012; Zhou et al., 2017).[10] More importantly, since our bound scales with the variance of $\widehat{Q}(\pi)$ for a single candidate policy π, this bound (6) can only be attained if $\widehat{Q}(\pi)$ is an efficient estimator of $Q(\pi)$. Thus, our bound establishes meaningful separation in terms of regret bounds available for different policy learners, and concretely establishes the relevance of the classic literature on semiparametrically efficient estimation (Bickel et al., 1998; Hahn, 1998; Hirano et al., 2003; Newey, 1994; Robins and Rotnitzky, 1995; Robins et al., 1995) to policy learning.

Our paper is structured as follows. First, in Section 2, we propose a method for policy learning, present a first regret bound of the form (6), and discuss implementation. Our main theoretical results are then developed in Section 3. Throughout our analysis, we prove uniform bounds that allow for both the data-generating distribution and the policy class Π to change with n; in particular, we allow the VC dimension of Π to grow as a positive power of n. Finally, Section 4 considers lower bounds for the minimax risk of policy learning, and shows that our regret bounds are optimal (up to constants) among all regret bounds that depend on the policy class Π via the VC dimension.

[8] The factor $1 + \log(V_{\max}/V_*)$ is perhaps unexpected, and may not be optimal. However, as discussed in Section 4, in many situations of interest we may expect to have $\log(V_{\max}/V_*) \approx 0$; and, in fact, our lower bounds are established for sequences of problems with $\log(V_{\max}/V_*) \to 0$. In these situations, the bound (6) becomes $R(\hat{\pi}) = \mathcal{O}_P(V_* \, \mathrm{VC}(\Pi)/n)$, which is the best possible bound of this form.

[9] The concrete algorithms we propose for policy learning—see Section 2.3 for details—most closely resemble methods developed by Dudík et al. (2011) and Zhang et al. (2012), who use doubly robust estimates for $\widehat{Q}(\pi)$ that are well-known to be semiparametrically efficient under appropriate conditions (Hahn, 1998; Robins and Rotnitzky, 1995). These papers, however, did not study the potential for efficiency gains from this method, and only focused on doubly robust consistency; moreover, the analytic tools needed to establish efficiency of these methods were not previously available in the literature. In particular, if one tried to apply the analysis of Kitagawa and Tetenov (2015) to the methods of Dudík et al. (2011) or Zhang et al. (2012) (or, in fact, to the method we propose in Section 2.3), one would still only obtain bounds of the form (4) that do not depend on the efficient variance V_*.

In this context, we also note the recent work of Zhou et al. (2017), who use a form of residualization to improve on outcome-weighted learning as proposed by Zhao et al. (2012), and find it to improve practical performance. However, the scoring method they use is inefficient whenever treatment propensities may deviate from $e(x) = 1/2$, and their theoretical analysis is not sharp enough to provide regret bounds that scale with the second moment of the scores.

[10] There is a related strand of literature in the machine learning community on "empirical Bernstein" regret bounds for weighted learning (Cortes et al., 2010; Maurer and Pontil, 2009; Swaminathan and Joachims, 2015). By applying these methods to our proposed policy learner, we could derive bounds of the type $R(\hat{\pi}) = \mathcal{O}_P(\sqrt{V_* \, \mathrm{VC}\left(\Pi\right) \log\left(n\right) / n})$. Much like our result, this bound has a conceptually pleasing quasi-optimal dependence on the variance of the efficient policy value estimate $\widehat{Q}(\pi^*)$ via V_*. However, unlike our result, this bound has an extraneous $\log(n)$ factor, which makes it inappropriate for asymptotic analysis; for large n, this bound is in fact worse than (4).

1.1 Interlude: Learning Simple Policies under Nonparametric Confounding

Our main result (6) is a hybrid parametric-nonparametric result. As in Kitagawa and Tetenov (2015), we assume that the practitioner wants to choose a policy π from among a "quasi-parametric" class of decision rules Π with finite VC-dimension (or, more generally, a finite entropy integral), but at the same only make non-parametric assumptions about the observational data that is used to learn this policy. Formally, we assume that the practitioner has access to n independent and identically distributed samples $(X_i,\, Y_i,\, W_i) \in \mathcal{X} \times \mathbb{R} \times \{\pm 1\}$ generated via potential outcomes $\{Y_i(\pm 1)\}$ such that $Y_i = Y_i(W_i)$. We require the treatment assignment to satisfy the identification condition (2); however, beyond that, we only assume generic regularity properties on

$$\mu_w(x) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right], \quad e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right], \tag{7}$$

and other aspects of the joint distribution of $(X_i,\, Y_i,\, W_i)$.

This juxtaposition of quasi-parametric and non-parametric setups may appear surprising at first glance; however, as also argued by Kitagawa and Tetenov (2015), we believe that separating assumptions for the nuisance components (8) and the class of policies Π is a necessary component of a comprehensive analysis of policy learning. On one hand, the nuisance components $\mu_w(x)$ and $e(x)$ are facts of nature the practitioner has no control over once the data has been collected, so assuming a non-parametric model for them seems like a prudent, conservative choice—especially given recent methodological developments that allow for these nuisance components to be estimated via powerful machine learning methods (Chernozhukov et al., 2016; van der Laan and Rose, 2011). On the other hand, the class of candidate policies Π is specified by the practitioner, and there may be many good reasons to place restrictions on it.

First, we may need to restrict the set of covariates that can be used by policies $\pi \in \Pi$. There are protected characteristics of people that may in principle affect the nuisance components $\mu_w(x)$ and $e(x)$, but cannot be used as decision variables: Kitagawa and Tetenov (2015) have an example where they measure age, gender and race, but do not use these features in choosing who should receive job training and/or job search assistance. It is also desirable that all features used by the candidate decision rules $\pi \in \Pi$ be reliably measured and available in a deployed system, and not be manipulable by participants.

Second, as discussed in Bhattacharya and Dupas (2012), we sometimes need to work with budget constraints that cap the total fraction of the population that may be treated. Furthermore, in some areas it may be desirable to pre-specify (or bound) treatment assignment rates by subgroup. For example, Kleinberg et al. (2017) study automated decision rules for mandating pre-trial detention, and emphasize a finding that they can substantially reduce predicted crime rates while maintaining fixed pre-trial detention rates across subgroups specified by race.

Third, in some application areas, it may be desirable for the policies $\pi \in \Pi$ to have a simple functional form, e.g., if they need to be audited or discussed by subject matter specialists, or if they need to be distributed in a non-electronic format. The formal distinction between the class of policy functions Π and our non-parametric model for the observational data provide a simple way to enforce all these desiderata (i.e., constraints on features, budget, or functional form) without making any problematic assumptions about the underlying distribution of $(X_i,\, Y_i,\, W_i)$.

There are of course some settings where there may be no meaningful difference between

the classes of functions used to estimate nuisance parameters and those used to estimate policies, and we are willing to learn semiparametric policy functions that are as complicated as our estimators for $(\mu_w(x),\, e(x))$. For example, when working with a single engineering system, e.g., a website wanting to target advertisements, we have access to a stable stream of incoming data and do not need to worry about changes in data availability or external validity. In cases like these, it makes sense to learn π using a bandit algorithm that can fit rich families of policy functions; see, e.g., Agarwal et al. (2014), Auer et al. (2002), Bastani and Bayati (2015), and references therein. However, in most public policy applications, we believe that using different classes of functions for $\{\pi(\cdot)\}$ versus $\{\mu_w(\cdot),\, e(\cdot)\}$ can be helpful, or even imperative.

2 From Efficient Policy Evaluation to Learning

Recall that we are interested in the following problem: We have n i.i.d. samples $(X_i,\, Y_i,\, W_i)$ drawn from a regular, unconfounded distribution satisfying (2), with

$$\mu_w^{(n)}(x) = \mathbb{E}_n\left[Y_i(w) \,\middle|\, X_i = x\right], \quad e^{(n)}(x) = \mathbb{P}_n\left[W_i = 1 \,\middle|\, X_i = x\right]. \tag{8}$$

Given such a data-generating distribution, we want to learn a policy assignment rule $\hat{\pi}_n : \mathcal{X} \to \{\pm 1\}$ belonging to a class $\hat{\pi}_n \in \Pi_n$, such as to make the regret (1) small. Following an extensive existing literature (Beygelzimer and Langford, 2009; Bottou et al., 2013; Chen et al., 2016; Dudík et al., 2011; Kitagawa and Tetenov, 2015; Swaminathan and Joachims, 2015; Zhao et al., 2012; Zhou et al., 2017), we focus on learners $\hat{\pi}_n$ obtained by optimizing a policy value estimate $\widehat{Q}_n(\pi)$ as in (3). Our goal is to find a class of efficient $\widehat{Q}_n$-estimators that yield $\hat{\pi}_n$-learners who inherit their efficiency properties.

In the previous paragraph—and in fact through the rest of the paper—we let both the data generating distribution for $(X_i,\, Y_i,\, W_i)$ and the policy class Π_n change with n; in particular, we will let the complexity of the class Π_n increase with n. All results will be uniform over a class of sequences of data generating distributions and policy classes satisfying regularity conditions discussed below. The only aspects of the problem we do not vary with n are the overlap bound η in (2) and a bound on the irreducible noise level $\mathrm{Var}_n\left[Y_i(w) \,\middle|\, X_i = x\right]$; this means that the efficient variance $V_*^{(n)}$ remains bounded as n gets large (although it may also change).

In this section, we start by specifying a concrete policy learning strategy below, and present a first regret bound for it in Section 2.2. Then, before presenting a proof, we discuss implementation of our method in Section 2.3, and show a simple example in Section 2.4.

2.1 Double Machine Learning for Policy Evaluation

Perhaps the simplest way to construct semiparametrically efficient estimators for $Q(\pi)$ is via doubly robust methods. These methods were originally studied by Hahn (1998) and Robins and Rotnitzky (1995) and, more recently, Belloni et al. (2017), Farrell (2015), van der Laan and Rose (2011), and others have considered extensions to high-dimensional settings. In this paper, we focus on doubly robust estimators obtained via the "double machine learning" construction advocated by Chernozhukov et al. (2016).

The core idea behind double machine learning is that, by relying on a modest amount of sample splitting, we can use machine learning methods to build treatment effect estimators that can be guaranteed to be efficient given only high-level conditions on the predictive

accuracy of the machine learning method. We build such estimators as follows: First divide the data into K evenly-sized folds and, for each fold $k = 1, ..., K$, run a machine learning estimator of our choice on the other $K-1$ data folds to estimate the functions $\mu^{(n)}_{\pm 1}(x)$ and $e^{(n)}(x)$; denote the resulting estimates $\hat{\mu}^{(-k)}_{\pm 1}(x)$ and $\hat{e}^{(-k)}(x)$ (with dependence on n suppressed). Then, given these pre-computed values, we estimate $Q_n(\pi)$ as

$$\widehat{Q}_{DML,n}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\widehat{\Gamma}_i, \quad \widehat{\Gamma}_i := \hat{\mu}^{(-k(i))}_{+1}(X_i) - \hat{\mu}^{(-k(i))}_{-1}(X_i) + W_i \frac{Y_i - \hat{\mu}^{(-k(i))}_{W_i}(X_i)}{\hat{e}^{(-k(i))}_{W_i}(X_i)}, \quad (9)$$

where $k(i) \in \{1, ..., K\}$ denotes the fold containing the i-th observation. Here, we have also used the short-hand

$$\hat{e}^{(-k(i))}_{W_i}(X_i) = \frac{1}{2} - W_i\left(\frac{1}{2} - \hat{e}^{(-k(i))}(X_i)\right) \quad (10)$$

to denote estimates of the class-specific propensities. The K-fold algorithmic structure used in (9) was proposed by Schick (1986) as a general purpose tool for efficient estimation in semiparametric models, and has also been used in Robins et al. (2008, 2017) and Wager et al. (2016).

Under weak assumptions, Chernozhukov et al. (2016) show that the double machine learning estimator (9) achieves the semiparametrically efficient rate for estimating $Q_n(\pi)$ (Hirano et al., 2003),[11]

$$\sqrt{n}\left(\widehat{Q}_{DML,n}(\pi) - Q_n(\pi)\right) \Big/ \sqrt{V^{(n)}(\pi)} \Rightarrow \mathcal{N}(0, 1), \quad (11)$$

$$V^{(n)}(\pi) = \mathrm{Var}_n\left[\pi(X)\tau^{(n)}(X)\right] + \mathbb{E}_n\left[\frac{\mathrm{Var}_n\left[Y(-1) \,\middle|\, X = X_i\right]}{1 - e^{(n)}(X_i)} + \frac{\mathrm{Var}_n\left[Y(+1) \,\middle|\, X = X_i\right]}{e^{(n)}(X_i)}\right],$$

provided the product of the root-mean squared error of the estimators $\hat{\mu}^{(n)}_{\pm 1}(x)$ and $\hat{e}^{(n)}(x)$ goes to zero faster than $1/\sqrt{n}$ (e.g., this would hold if both estimators were $o_P(n^{1/4})$-consistent). For a review of conditions under which such convergence is possible, see Chernozhukov et al. (2016); for example, it is enough that $e^{(n)}(\cdot)$ and $\mu^{(n)}_w(\cdot)$ belong to appropriate L_2-Sobolev classes. This estimator is closely related to the classical semiparametric two-stage methods studied by, e.g., Hahn (1998), Newey (1994) and Robins and Rotnitzky (1995).

Throughout our analysis, we will make the following assumption about the machine learning method underlying (9).

Assumption 1 (Consistent machine learning)**.** Whenever we use a double machine learning estimator $\widehat{Q}_{DML,n}$ constructed as in (9), we assume that the machine learning methods used to construct our estimator satisfy the following consistency guarantees. The methods must be uniformly consistent,

$$\sup_{x\in\mathcal{X}}\left|\hat{\mu}^{(n)}_w(x) - \mu^{(n)}_w(x)\right|, \quad \sup_{x\in\mathcal{X}}\left|\hat{e}^{(n)}_{\pm 1}(x) - e^{(n)}_{\pm 1}(x)\right| \to_p 0. \quad (12)$$

Moreover, the product of the L_2-errors of both methods must converge as $n^{1/2}$:

$$\mathbb{E}\left[\left(\hat{\mu}^{(n)}_w(X) - \mu^{(n)}_w(X)\right)^2\right] \leq \frac{a(n)}{n^{\zeta_\mu}}, \quad \mathbb{E}\left[\left(1/\hat{e}^{(n)}_{\pm 1}(X) - 1/e^{(n)}_{\pm 1}(X)\right)^2\right] \leq \frac{a(n)}{n^{\zeta_e}}, \quad (13)$$

[11] Recall that we write the conditional average treatment effect as $\tau^{(n)}(x) = \mu^{(n)}_{+1}(x) - \mu^{(n)}_{-1}(x)$.

for some constants $0 < \zeta_\mu, \zeta_e < 1$ with $\zeta_\mu + \zeta_e \geq 1$, and some sequence $a(n) \to 0$. Here X is taken to be an independent test example drawn from the same distribution as the training data.[12]

Given these conditions, the results of Chernozhukov et al. (2016) imply that, for any single policy π, double machine learning policy evaluators $\widehat{Q}_{DML}(\pi)$ built via methods satisfying Assumption 1 are asymptotically efficient for estimating $Q(\pi)$. In Section 3.3, we extend their analysis and establish conditions under which such convergence holds uniformly over the whole class $\pi \in \Pi$; this result will then play a key role in establishing strong regret bounds for policy learning. Before presenting the result in more detail, however, we first discuss the type of regret bound it enables, and how to implement a concrete policy learner building on our choice of $\widehat{Q}_{DML}(\pi)$.

2.2 A Motivating Result

To get a feeling for the types of results we can obtain for policy learning with double machine learning, we consider below the case where Π belongs to a VC class. This setup allows us to state a result without resorting to too much notation. As discussed in the introduction, our results will also depend on the following upper bound for the worst-case efficient variance for estimating any policy:

$$V_{\max}^{(n)} = \mathbb{E}_n\left[\left(\tau^{(n)}(X)\right)^2\right] + \mathbb{E}_n\left[\frac{\mathrm{Var}_n\left[Y(-1)\,\middle|\,X = X_i\right]}{1 - e^{(n)}(X_i)} + \frac{\mathrm{Var}_n\left[Y(+1)\,\middle|\,X = X_i\right]}{e^{(n)}(X_i)}\right], \quad (14)$$

and note that $V^{(n)}(\pi) = V_{\max}^{(n)} - Q_n^2(\pi)$ for any policy π. We develop the technical tools required to prove this result in Section 3.

Theorem 1. *Define $\widehat{Q}_{DML,n}(\pi)$ as in (9), and let $\hat{\pi}_n = \mathrm{argmin}_{\pi\in\Pi_n} \widehat{Q}_{DML,n}(\pi)$. Given unconfoundedness and overlap (2) and Assumption 1, suppose moreover that the irreducible noise $\varepsilon_i = Y_i - \mathbb{E}_n\left[Y_i \,\middle|\, X_i, W_i\right]$ is both uniformly sub-Gaussian conditionally on X_i and W_i and has second moments uniformly bounded from below, $\mathrm{Var}\left[\varepsilon_i \,\middle|\, X_i = x, W_i = w\right] \geq s^2$, and that the conditional average treatment effect $\tau^{(n)}(x)$ is uniformly bounded in x and n. Finally, suppose that Π_n is a VC class of dimension $\mathrm{VC}(\Pi_n)$ bounded by*

$$\mathrm{VC}(\Pi_n) = \mathcal{O}\left(n^\beta\right), \quad \beta \leq \min\left\{\zeta_\mu, \zeta_e\right\}, \quad \beta < 1/2. \quad (15)$$

Then, for any $\delta > 0$, there is a universal constant[13] C_δ, as well as a threshold N that depends on the constants used to define the regularity assumptions such that, with probability at least $1 - \delta$,

$$R_n(\hat{\pi}_n) \leq C_\delta \sqrt{\mathrm{VC}(\Pi_n)\, V^{(n)}(\pi^*) \left(1 + \log\left(\frac{V_{\max}^{(n)}}{V^{(n)}(\pi^*)}\right)\right) \Big/ n}, \text{ for all } n \geq N, \quad (16)$$

[12] A notable special case of this assumption is when $\zeta_\mu = \zeta_e = 1/2$; 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 propensity 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 μ-model with accuracy of the e-model, provided the product of the error rates is controlled (Farrell, 2015).

[13] Throughout this paper, we will use C and C_δ to denote different universal constants; no two instantiations of C and C_δ should be assumed to denote the same constant.

where $R_n(\cdot)$ denotes regret for the n-th data-generating distribution, $V^{(n)}(\pi)$ denotes the semiparametric efficient variance for policy evaluation (11) and $V^{(n)}_{\max}$ is as defined in (14).

2.3 Implementation via Weighted Classification

In the previous sections, we established regret bounds for the policy $\hat{\pi}_{DML}$ obtained by maximizing $\widehat{Q}_{DML}(\pi)$ as defined in (9). In order to carry out this optimization, we follow Beygelzimer and Langford (2009), Kitagawa and Tetenov (2015), Zhang et al. (2012), Zhao et al. (2012) and Zhou et al. (2017), and note that $\hat{\pi}_{DML}$ can also be understood as the empirical risk minimizer in a weighted classification problem:

$$\hat{\pi} = \text{argmax}_{\pi \in \Pi} \left\{ \frac{1}{n} \sum_{i=1}^{n} \lambda_i Z_i \pi(X_i) \right\}, \quad \lambda_i = \left| \widehat{\Gamma}_i \right|, \quad Z_i = \text{sign} \left(\widehat{\Gamma}_i \right), \tag{17}$$

i.e., we want to train a classifier with response Z_i using weights λ_i; recall that $\widehat{\Gamma}_i$ was defined in (9). Given this formulation as a weighted classification problem, we can use standard off-the-shelf tools for weighted classification to learn $\hat{\pi}$, e.g., classification trees (Breiman et al., 1984), support vector machines (Cortes and Vapnik, 1995), or best-subset empirical risk minimization (Chen and Lee, 2016; Greenshtein et al., 2006).[14]

Several other proposals also fit into this framework, with different choices of $\widehat{\Gamma}_i$. Zhao et al. (2012) assume a randomized controlled trial and use $\widehat{\Gamma}_i = W_i \, Y_i \, / \, \mathbb{P}\left[W_i = 1\right]$, while Kitagawa and Tetenov (2015) use inverse-propensity weighting $\widehat{\Gamma}_i = W_i \, Y_i \, / \, \hat{e}_{W_i}(X_i)$. In an attempt to stabilize the weights, Beygelzimer and Langford (2009) introduce an "offset"

$$\widehat{\Gamma}_i = \frac{W_i}{\hat{e}_{W_i}(X_i)} \left(Y_i - \frac{\max\left\{Y_i\right\} + \min\left\{Y_i\right\}}{2} \right),$$

while Zhou et al. (2017) go further and advocate

$$\widehat{\Gamma}_i = \frac{W_i}{\hat{e}_{W_i}(X_i)} \left(Y_i - \frac{\hat{\mu}_{+1}(X_i) + \hat{\mu}_{-1}(X_i)}{2} \right).$$

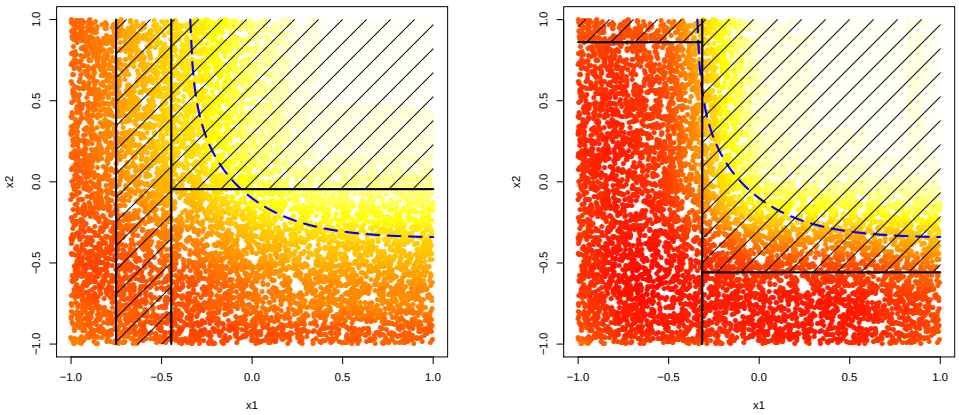
None of the above methods, however, are built on semiparametrically efficient policy evaluation, and so they do not fit into the class of algorithms covered by Theorem 1. Finally, the method advocated by Zhang et al. (2012) actually takes the same form as our procedure (17). However, the paper by Zhang et al. (2012) does not provide regret bounds; moreover, they do not use "cross-fitting" or "cross-estimation" as in (9), so it is unclear under what conditions their method satisfies the bounds from Theorem 1.[15]

2.4 A Simple Illustration

We illustrate our approach with a simple simulation example. Suppose that we have access to data $(X, Y, W) \in [-1, 1]^2 \times \mathbb{R} \times \{\pm 1\}$, and want to learn a policy function $\pi(\cdot)$ that

[14] In our discussion so far, we have largely left aside questions on how to estimate $\hat{\mu}_w(\cdot)$ and $\hat{e}(\cdot)$, provided that they are obtained using some machine learning method that satisfies Assumption 1. However, in terms of finite-sample performance, experience suggests that it may be preferable to use methods for $\hat{\mu}_{+1}(\cdot)$ and $\hat{\mu}_{-1}(\cdot)$ that estimate both arms simultaneously while explicitly seeking out treatment effect heterogeneity (Athey et al., 2016b; Imai and Ratkovic, 2013).

[15] Other related methods, such as those by Dudík et al. (2011) and Swaminathan and Joachims (2015), apply in a setting with multiple available treatments and so do not directly fit into the setting of (17). Extending our efficiency analysis to the multi-treatment regime would be of considerable interest.



Inverse-propensity weighting Double machine learning

Figure 1: Results from two attempts at learning a policy π by counterfactual risk minimization (3) over depth-2 decision trees, as described in Section 2.4, with $\widehat{Q}$-estimators obtained by both inverse-propensity weighting and double machine learning. The dashed blue line denotes the Bayes-optimal decision rule (treat in the upper-right corner, do not treat elsewhere); the solid black lines denote learned policies π (treat in the shaded regions, do not treat elsewhere). We also use a heat map to depict the average policies learned across 200 simulations; the darkest red regions are never treated and lightest yellow regions are always treated. Simulations were performed using $n = 500$ samples drawn according to (18).

can be written as a depth-2 decision tree. Here, the data is independent and identically distributed as

$$X_i \sim U\left([-1,\,1]^2\right),\;\; W_i \,\big|\, X_i \sim \text{Bern}\left(e(x)\right),\;\; Y_i \,\big|\, W_i,\, X_i \sim \mathcal{N}\left(\mu(X_i) + \tau(X_i)W_i/2,\, 1\right), \quad (18)$$

with $e(x) = 1/(1+e^{-(x_1+x_2)})$, $\mu(x) = 2e^{-(x_1+x_2)}$, and $\tau(x) = 2/[(1+e^{-4x_1})(1+e^{-4x_2})]-0.4$. For our purposes, the salient facts about this data-generating distribution are that it is unconfounded (2); however, it cannot be represented by trees, and the Bayes-optimal policy is not in our class Π of interest, i.e., depth-2 decision trees.

In Figure 1, we show results for learning π using two different counterfactual risk minimization strategies of the form (3), but with different $\widehat{Q}$-estimators. The left panel obtains $\widehat{Q}$ by inverse-propensity weighting; conversely, the right panel uses an efficient double machine learning $\widehat{Q}$-estimator. We see that the policies $\hat{\pi}$ learned via efficient policy evaluation are much better than those learned by inverse-propensity weighting. Across 200 simulation runs, inverse-propensity weighting led to a mean regret of 0.143, whereas double machine learning got a mean regret of 0.063 relative to the best possible depth-2 tree. Figure 1 shows both a single realization of each method, as well as the average policy learned across 200 simulation runs.

In terms of specifics, we started by learning $\hat{\mu}_{\pm 1}(\cdot)$ and $\hat{e}(\cdot)$ via a lasso (Tibshirani, 1996) on a polynomial basis expansion (with interactions), all while using out-of-fold prediction as in (9). For $\hat{\mu}_{\pm 1}(\cdot)$, we fit both response functions simultaneously while writing them in terms of a main effect $(\hat{\mu}_{+1}(\cdot) + \hat{\mu}_{-1}(\cdot))/2$ and a treatment effect $(\hat{\mu}_{+1}(\cdot) - \hat{\mu}_{-1}(\cdot))/2$; when

the treatment effect is weaker than the main effect, this re-parametrization can interact with the lasso penalty in a way that improves the resulting fit (Imai and Ratkovic, 2013). Finally, we got $\hat{\pi}$ by optimizing (17) over depth-2 trees with the `R`-package `evtree`, which learns an optimal classification tree using an evolutionary algorithm (Grubinger et al., 2014). Inverse-propensity weighting uses $\widehat{\Gamma}_i = Y_i / \hat{e}_{W_i}(X_i)$, while our method uses the form in (9).

3 Theoretical Development

As is common in the literature on semiparametric estimation, our proof is built around a study of the efficient influence function for policy evaluation,

$$\widetilde{Q}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\Gamma_i^{(n)}, \quad \Gamma_i^{(n)} := \mu_{+1}^{(n)}(X_i) - \mu_{-1}^{(n)}(X_i) + W_i \frac{Y_i - \mu_{W_i}^{(n)}(X_i)}{e_{W_i}^{(n)}(X_i)}. \tag{19}$$

Note that $\widetilde{Q}_n(\pi)$ can be understood as a version of $\widehat{Q}_n(\pi)$ computed by an oracle who has access to the true functions $\mu_w^{(n)}(x)$ and $e^{(n)}(x)$. Many classical results in semiparametric efficiency theory rely on characterizing the realizable estimators $\widehat{Q}_n(\pi)$ in terms of the oracle quantities $\widetilde{Q}_n(\pi)$: Classical results going back at least to Newey (1994) show that, under appropriate regularity conditions, two-stage estimators of the form (9) are asymptotically equivalent to an average of the efficient influence function,

$$\sqrt{n}\left(\widehat{Q}_n(\pi) - \widetilde{Q}_n(\pi)\right) \to_p 0. \tag{20}$$

This type of idea approach is also standard in the literature on treatment effect estimation (e.g., Chernozhukov et al., 2016; Hahn, 1998; Robins and Rotnitzky, 1995).

Our proof is structured as follows. Having spelled out assumptions about the policy class below, we proceed in Section 3.2 to prove a regret bound that would be available to an analyst who could optimize the infeasible objective $\widetilde{Q}_n(\pi)$ rather then the feasible double machine learning objective $\widehat{Q}_n(\pi)$. Then in Section 3.3, we establish a strengthening of (20) that holds uniformly over all $\pi \in \Pi$, and use this coupling to establish a first result about learning with $\widehat{Q}_n(\pi)$. Finally, in Section 3.4, we re-visit and strengthen our bounds under stronger assumptions on Π_n that hold, for example, when Π_n is a VC-class.

3.1 Assumptions about the Policy Class

Although we stated our first result, Theorem 1, under the simple assumption that Π is a Vapnik-Chervonenkis class, we will develop our technical results under weaker, more abstract assumptions on Π. In order to obtain regret bounds as in (16) that decay as $1/\sqrt{n}$, we of course need some control over the complexity of the class Π.

Here, we do so using bounds on the Hamming entropy of Π. For any discrete set of points $\{X_1, ..., X_m\}$ and any $\varepsilon > 0$, define the ε-Hamming covering number $N_H(\varepsilon, \Pi, \{X_1, ..., X_m\})$ as the smallest number of policies $\pi : \{X_1, ..., X_m\} \to \{\pm 1\}$ (not necessarily contained in Π) required to ε-cover Π under Hamming distance,

$$H(\pi_1, \pi_2) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\left(\{\pi_1(X_j) \neq \pi_2(X_j)\}\right).$$

Then, define the ε-Hamming entropy of Π as $\log(N_H(\varepsilon, \Pi))$, where

$$N_H(\varepsilon, \Pi) = \sup\{N_H(\varepsilon, \Pi, \{X_1, ..., X_m\}) : X_1, ..., X_m \in \mathcal{X}; m \geq 1\}$$

is the number of functions needed to ε-cover Π under Hamming distance for any discrete set of points. We note that this notion of entropy is purely geometric, and does not depend on the distribution used to generate the X_i. Finally, we assume the following:

Assumption 2 (Entropy bound)**.** We assume that there are constants C, $\beta \geq 0$ and $\omega > 0$ such that $\beta + \omega < 1/2$, and that the Hamming entropy of Π_n is bounded by

$$\log(N_H(\varepsilon, \Pi_n)) \leq Cn^{\beta}\varepsilon^{-\omega} \quad \text{for all} \quad 0 < \varepsilon < 1 \text{ and } n \in \mathbb{N}. \tag{21}$$

Given this assumption define the complexity κ of the class Π_n in terms of a variant of the classical entropy integral of Dudley (1967):[16]

$$\kappa(\Pi_n) = \int_0^1 \sqrt{\log(N_H(\varepsilon^2, \Pi_n))}\, d\varepsilon. \tag{22}$$

Using entropy integrals to bound model class complexity is ubiquitous in empirical process theory (see, e.g., Boucheron et al. (2013)), and easily allows us to specialize to more restrictive cases. In particular, in one example of particular interest, it is well known that if Π is a Vapnik-Chervonenkis class, then (Haussler, 1995)

$$\log(N_H(\varepsilon, \Pi)) \leq d\left(\log\left(\varepsilon^{-1}\right) + \log(2) + 1\right) + \log(d+1) + 1, \text{ with } d := \text{VC}(\Pi). \tag{23}$$

Thus, we immediately see that Assumption 2 holds whenever $\Pi_n = \Pi$ is a fixed VC class, and can use (23) to verify that $\kappa^2(\Pi) \leq 6d$ for any value of $d = 1, 2, ...$ Moreover, when Π_n is a sequence of VC classes of increasing dimension d_n, Assumption 2 still holds whenever $d_n = \mathcal{O}\left(n^{\beta}\right)$ for some $\beta < 1/2$ (with, e.g., $\omega = (1/2 - \beta)/2$).

There has also been some recent interest in developing tree-based decision rules (Athey and Imbens, 2016; Kallus, 2017; Su et al., 2009). If we let Π consist of the set of all depth-L decision trees with $X_i \in \mathbb{R}^d$, we can verify that[17]

$$\log(N_H(\varepsilon, \Pi)) = \mathcal{O}\left(2^L \log\left(\varepsilon^{-1}\right) + 2^L \log(d) + L2^L\right). \tag{24}$$

Then, letting the depth L_n grow as $L_n = \lfloor \beta \log_2(n) \rfloor$ for some $\beta < 1$ Assumption 2 again holds with $\omega = (1/2 - \beta)/2$.

Finally, we note that further high-level constraints on Π as discussed in Section 1.1, e.g., budget constraints or constraints on marginal treatment rates among subgroups, simply reduce the complexity of the policy class Π and thus do not interfere with the present assumptions.

[16] Assumption 2 immediately guarantees that this integral is finite.

[17] To establish this result for trees, one can follow Bartlett and Mendelson (2002) and view each tree-leaf as a conjunction of L boolean functions, along with a sign. A simple argument then shows that a library of $4d^2L2^L\varepsilon^{-1}$ boolean functions lets us approximate each leaf to within Hamming error $2^{-L}\varepsilon$; and so we can also also approximate the tree to within ε Hamming error. The resulting bound on $N_H(\varepsilon, \Pi)$ follows by noting that a full tree has $2^L - 1$ splits, and so can be approximated using $2^L - 1$ of these boolean functions.

3.2 Rademacher Complexities and Oracle Regret Bounds

We start our analysis by characterizing the regret of an oracle learner who has access to the functions $\mu_{\pm 1}^{(n)}(\cdot)$ and $e^{(n)}(\cdot)$, and chooses their policy $\hat{\pi}_n$ by optimizing the infeasible value estimator $\widetilde{Q}_n(\pi)$ as defined in (19). The advantage of studying this oracle is that it allows us, for the time being, to abstract away from the specific machine learning methods used to obtain $\widehat{Q}_n(\pi)$, and instead to focus on the complexity of counterfactual risk minimization over the class Π_n.

Specifically, our present goal is to study concentration of the empirical process $\widetilde{Q}_n(\pi) - Q_n(\pi)$ for all $\pi \in \Pi_n$. Recalling the definition of $\Gamma_i^{(n)}$ from (19), a convenient way to bound the supremum of our empirical process of interest is by controlling its Rademacher complexity $\mathcal{R}_n(\Pi_n)$, defined as

$$\mathcal{R}_n(\Pi_n) = \sup_{\pi \in \Pi_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i \Gamma_i^{(n)} \pi(X_i) \right\} \tag{25}$$

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

Following this proof strategy, we start by providing a bound for $\mathcal{R}_n(\Pi_n)$ that scales as $\sqrt{\mathbb{E}\left[\Gamma^2\right]/n}$. Despite its simple form, we are not aware of existing proofs of such results in the literature. Bounds that scale as $\max\{\Gamma_i\}/\sqrt{n}$ are standard but, in our setting, are not strong enough to move past results of the type (4) as obtained by, e.g., Kitagawa and Tetenov (2015). Meanwhile, bounds that scale as $\sqrt{\mathbb{E}\left[\Gamma^2\right]\log(n)/n}$ are developed by Cortes et al. (2010) and Maurer and Pontil (2009); however, the additional $\log(n)$ factor makes these bounds inappropriate for asymptotic analysis.

Lemma 2. *Suppose that the class Π_n satisfies Assumption 2, and that the weights $\Gamma_i^{(n)}$ in (25) are drawn from a sequence of uniformly sub-Gaussian distributions with variance bounded from below,*[18]

$$\mathbb{P}\left[\left|\Gamma_i^{(n)}\right| > t\right] \leq C_\nu \, e^{-\nu t^2} \text{ for all } t > 0, \quad \text{Var}\left[\Gamma_i^{(n)}\right] \geq s^2, \tag{26}$$

for some constants C_ν, ν, $s > 0$ that do not depend on n. Then, there is a universal constant $C > 0$ for which

$$\mathbb{E}\left[\mathcal{R}_n\left(\Pi_n\right)\right] \leq 8\left(\kappa\left(\Pi_n\right) + C\right) \sqrt{\mathbb{E}\left[\left(\Gamma_i^{(n)}\right)^2\right] \Big/ n} + \mathcal{O}\left(\frac{\sqrt{\log(n)}}{n}\right), \tag{27}$$

where $\kappa(\Pi_n)$ is the complexity of Π_n as defined in (22).

Given this Rademacher complexity bound, we can obtain a uniform concentration bound for $\widetilde{Q}(\pi)$ using standard arguments. Here, we refine an argument of Bartlett and Mendelson (2002) using Talagrand's inequality to obtain a bound that scales as $\sqrt{\mathbb{E}\left[\Gamma_i^2\right]}$ rather than $\sup|\Gamma_i|$. In the statement of the result below, we note that $V_{\max} = \mathbb{E}\left[\Gamma_i^2\right]$, as is clear from (14).

[18] Due to typographical concerns, we will frequently omit the superscript (n) in the body of the text when there is no risk of confusion.

Theorem 3. *Under the conditions of Lemma 2, the averaged efficient influence functions $\widetilde{Q}_n(\pi)$ concentrate uniformly: There is a universal constant $C > 0$ such that, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{\pi \in \Pi_n} \left| \widetilde{Q}_n(\pi) - Q_n(\pi) \right| \leq \left(16\kappa(\Pi_n) + \sqrt{2\log(\delta^{-1})} + C \right) \sqrt{V_{\max}^{(n)} / n} + \mathcal{O}\left(\log(n)\kappa(\Pi_n) / n \right). \tag{28}$$

If we set $\tilde{\pi}_n \in \operatorname{argmax}\{\widetilde{Q}_n(\pi) : \pi \in \Pi\}$, then again with probability at least $1 - \delta$,

$$R_n(\tilde{\pi}_n) \leq 2 \left(16\kappa(\Pi_n) + \sqrt{2\log(\delta^{-1})} + C \right) \sqrt{\frac{V_{\max}^{(n)}}{n}} + \mathcal{O}\left(\frac{\log(n)\kappa(\Pi_n)}{n} \right), \tag{29}$$

where $R_n(\tilde{\pi}_n)$ stands for the regret of policy $\tilde{\pi}_n$.

3.3 Uniform Coupling with the Efficient Score

In the previous section, we established risk bounds that would hold if we could optimize the infeasible value function $\widetilde{Q}_n(\pi)$; 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 double machine learning estimator $\widehat{Q}_n(\pi) = \widehat{Q}_{DML,n}(\pi)$. For a single, fixed policy π, Chernozhukov et al. (2016) showed that $\widetilde{Q}_n(\pi)$ and $\widehat{Q}_n(\pi)$ are asymptotically equivalent, 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 pre-determined 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 (20) that holds simultaneously for all $\pi \in \Pi_n$.

Lemma 4. *Under the conditions of Lemma 2, suppose that we obtain $\widehat{Q}_n(\pi) = \widehat{Q}_{DML,n}(\pi)$ by double machine learning according to Assumption 1 and moreover that overlap holds as in (2). Then*

$$\begin{aligned} &\sqrt{n} \sup \left\{ \left| \widehat{Q}_n(\pi) - \widetilde{Q}(\pi) \right| : \pi \in \Pi_n \right\} \Big/ a\left(\left(1 - K^{-1}\right) n \right) \\ &\quad = \mathcal{O}_P\left(1,\ \kappa(\Pi_n) \,/\, \sqrt{n^{\min\{\zeta_\mu, \zeta_e\}}} \right), \end{aligned} \tag{30}$$

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

The above result is perhaps surprisingly strong: Provided that the complexity of Π_n, $\kappa(\Pi_n)$, 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 complexity $\kappa(\Pi_n)$ of the class Π_n does not affect the leading-order constants in the bound. In other words, in terms of the coupling of $\widetilde{Q}_n(\pi)$ and $\widehat{Q}_n(\pi)$, 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 by double machine learning is not the only way to construct efficient estimators for the value of a single policy π—for example, Hirano et al. (2003) show that inverse-propensity weighting with non-parametrically estimated propensity scores may also be efficient—but it plays a key role in the proof of Lemma 4. It is far from obvious that other efficient methods

for evaluating a single policy π, such as that of Hirano et al. (2003), would lead to equally strong uniform couplings over the whole class Π_n.

Given our coupling lemma, the following result is an immediate corollary of Theorem 3. In terms of assumptions, we note that assuming overlap as in (2), sub-Gaussianity of the irreducible noise $\varepsilon_i = Y_i - \mathbb{E}_n\left[Y_i \,\middle|\, X_i, W_i\right]$, and uniform boundedness of the conditional average treatment effect $\tau^{(n)}(x)$ lets us guarantee that the weights $\Gamma_i^{(n)}$ in (25) are sub-Gaussian; meanwhile, the lower bound on $\mathrm{Var}_n\left[\varepsilon_i\right]$ induces a lower bound on $\mathbb{E}_n\left[\Gamma_i^{(n)2}\right]$.

Theorem 5. *Suppose that Assumption 1 and 2 hold, that we have unconfoundedness and overlap as in (2), and that the irreducible noise $\varepsilon_i = Y_i - \mathbb{E}\left[Y_i \,\middle|\, X_i, W_i\right]$ is both uniformly sub-Gaussian conditionally on X_i and W_i and has second moments uniformly bounded from below, $\mathrm{Var}\left[\varepsilon_i \,\middle|\, X_i = x, W_i = w\right] \geq s^2$, and that $\tau^{(n)}(x)$ is uniformly bounded in x and n. Suppose, moreover, that the candidate policy class Π_n grows slowly enough that $\limsup_n n^{-\min\{\zeta_\mu, \zeta_e\}} \kappa^2(\Pi_n) < \infty$. Then, for any $\delta > 0$, there is a universal constant C_δ and a problem-specific threshold N such that*

$$R_n(\hat{\pi}_n) \leq C_\delta \max\left\{\kappa(\Pi_n), 1\right\} \sqrt{V_{\max}^{(n)} / n} \tag{31}$$

with probability at least $1-\delta$ for all $n \geq N$, where $\hat{\pi}_n$ optimizes the double machine learning risk estimate as in (3).

The above bound is close to our desideratum: It obtains regret bounds for a realizable policy that have the desired dependence on $\kappa(\Pi_n)$ and n (recall that, for VC-classes, $\kappa(\Pi_n) = \mathcal{O}(\sqrt{\mathrm{VC}(\Pi_n)})$), and scales with a semiparametric variance bound rather than with a crude bound on, e.g., $\sup|Y_i|$. However, a down-side of the bound (31) is that it depends on $V_{\max}^{(n)}$, i.e., a worst-case bound for the semiparametric efficient variance for evaluating any policy, and not as $V_*^{(n)}$, the semiparametric efficient variance for evaluating the optimal policy π^*. In the following section, we seek to replace the dependence on $V_{\max}^{(n)}$ with one on $V_*^{(n)}$, at the expense of stronger assumptions on the class Π_n.

3.4 Improved Bounds via Slicing

To move past the $V_{\max}^{(n)}$ scaling above, we need a closer analysis that cuts the space Π_n into strata of policies that have comparable values of $V^{(n)}(\pi)$, and then develop concentration bounds for these strata separately. This "slicing" idea is common in the literature, and has been used in different contexts by, e.g., Bartlett et al. (2005) and Giné and Koltchinskii (2006).

The reason we might expect slicing to work in our case is that, as discussed earlier, the efficient variance for evaluating any given policy π is $V^{(n)}(\pi) = V_{\max}^{(n)} - Q_n^2(\pi)$. Thus, any "good" policy, i.e., with a large value $Q_n(\pi)$, must also have a small efficient variance $V^{(n)}(\pi)$. More specifically, letting $\Pi_{\lambda,n}$ denote the set of policies with with regret at most λ,

$$\Pi_{\lambda,n} = \left\{\pi \in \Pi_n : R(\pi) \leq \lambda\right\}, \tag{32}$$

we immediately see that

$$\sup\left\{V^{(n)}(\pi) : \pi \in \Pi_{\lambda,n}\right\} \geq V_\lambda^{(n)} := V^{(n)}(\pi_n^*) + 2\lambda Q_n(\pi_n^*), \tag{33}$$

where π_n^* is an optimal policy.

This improved variance bound suggests an argument proceeding in two stages: First, Theorem 5 already established that the learned policy $\hat{\pi}_n$ has regret going to 0 and so $\mathbb{P}[\hat{\pi} \in \Pi_{\lambda,n}] \to 1$ for any fixed $\lambda > 0$; then, in a second stage, we use the improved variance bounds in (33) to get tighter concentration bounds for $\hat{\pi}_n$.

The fact that such a slicing argument works is not to be taken for granted; and, in our case, is a property that hinges crucially on the fact that, if we use an efficient method for policy evaluation, then there do not exist any policies π_n that simultaneously have low regret and are hard to evaluate (i.e., $V^{(n)}(\pi_n)$ is large). Given other evaluation methods, e.g., inverse propensity weighting as used in Kitagawa and Tetenov (2015) or Swaminathan and Joachims (2015), such a slicing argument may not work.

Below we establish a concentration bound for $\Pi_{\lambda,n}$ under the following assumption on the entropy of Π_n: for some $\alpha_n > 0$,

$$\log\left(N_H\left(\varepsilon,\, \Pi_n\right)\right) \leq \alpha_n \log\left(\varepsilon^{-1}\right), \text{ for all } 0 < \varepsilon < \frac{1}{2}. \tag{34}$$

Note that if Π_n is a VC-class then the above holds with $\alpha_n \leq 6\,\mathrm{VC}(\Pi_n)$; see (23).

Theorem 6. *Under the conditions of Theorem 5, suppose moreover that (34) holds for a sequence $\alpha_n = \mathcal{O}\left(n^{\beta}\right)$ for some $\beta < 1/2$, and let $\lambda > 0$ be predetermined. Then, for any $\delta > 0$ there is a universal constant C_δ, as well as a potentially problem-specific threshold N, such that for all $n \geq N$, with probability at least $1 - \delta$,*

$$\sup_{\pi \in \Pi_{\lambda,n}} \left|\widetilde{Q}_n\left(\pi\right) - Q_n\left(\pi\right)\right| \leq C_\delta \sqrt{\frac{\alpha_n}{n} V_\lambda^{(n)} \left(1 + \log\left(\frac{V_{\max}^{(n)}}{V_\lambda^{(n)}}\right)\right)}, \tag{35}$$

where α_n controls the complexity of Π_n via (34).

Although the above bound may superficially look like a direct extrapolation from (28), we caution that the proof relies on a subtly different construction than that used in Lemma 2, requiring stronger assumptions. In particular, the additional factor $\log(V_{\max}^{(n)}/V_*^{(n)})$ in the bound below is directly tied to the entropy growth rate assumed in (34); and in fact is closely related to the $\log(n)$ factor appearing in the empirical Bernstein bounds of Cortes et al. (2010) and Maurer and Pontil (2009).

Given this result, the proof of our main result follows immediately. Note that, whenever the condition (34) holds, we can replace the term $\mathrm{VC}(\Pi_n)$ with $\alpha_n/6$ in the statement of Theorem 1.

Proof of Theorem 1. Given that Π_n is a VC class, recall that (34) holds with $\alpha_n \leq 6\,\mathrm{VC}(\Pi_n)$. Now, set

$$\lambda = \frac{1}{3} \liminf_{n\to\infty} \left\{V^{(n)}(\pi^*)\right\} \Big/ \limsup_{n\to\infty} \left\{Q_n(\pi^*)\right\},$$

so $V_\lambda^{(n)} \leq 2V_*^{(n)}$ for large enough n. By Theorem 5, we know that $\mathbb{P}[\hat{\pi} \in \Pi_\lambda] \to 1$, and so (16) follows immediately from Theorem 6 paired with Lemma 4 (recall that, given (34), $\kappa^2(\Pi_n) = \mathcal{O}\left(\alpha_n\right)$). □

4 A Lower Bound for Minimax Policy Regret

To complement the upper bounds given in Theorems 1 and 5, we also present lower bounds on the minimax risk for policy learning, with the goal of showing that these upper bounds

are optimal up to constants. Of course, any optimality statement about upper bounds must be considered with care, as it is sensitive to the class of bounds under consideration. For example, after proving the bound (4), Kitagawa and Tetenov (2015) effectively argue that their bound is optimal—and, in fact, it is the best possible regret bound for policy learning that only depends on M, η, $\text{VC}(\Pi)$ and n, because sometimes M/η is a sharp bound for the semiparametric variance V_* (again, up to constants). But in this paper, we found that it is possible to meaningfully improve on the bound of Kitagawa and Tetenov (2015) if we are willing to have a more nuanced dependence on the $\{X_i, Y_i, W_i\}$-distribution.

In this light, our goal is to show that our bounds are the best possible regret bounds that flexibly account for the joint distribution of $\{X_i, Y_i, W_i\}$, but only depend on the policy class Π through the Vapnik-Chervonenkis dimension $\text{VC}(\Pi)$. This approach is in line with existing results in the machine learning literature: It is well known that regret bounds for empirical risk minimization over Π based on structural summaries of Π (such as the VC dimension) may sometimes 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.

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 $m(x)$ 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

$$\begin{aligned} &X_i \sim \text{Uniform}\left(\mathcal{X}_s\right), \quad W_i \,\big|\, X_i \sim 2 \cdot \text{Bernoulli}(e(X_i)) - 1, \\ &Y_i \,\big|\, X_i, W_i \sim \mathcal{N}\left(m(X_i) + \left(\frac{W_i + 1}{2} - e(X_i)\right) \frac{\tau(X_i)}{\sqrt{n}}, \sigma^2(X_i)\right). \end{aligned} \tag{36}$$

Because of the number of derivatives assumed on $m(x)$ and $e(x)$, it is well known that simple series estimators satisfy Assumption 1,[19] and so Theorem 1 immediately implies that, under unconfoundedness,

$$R_n\left(\hat{\pi}_n\right) = \mathcal{O}_P\left(\sqrt{\frac{V_* \,\text{VC}\left(\Pi\right)}{n}}\right), \quad V_* = \mathbb{E}\left[\frac{\sigma^2(X_i)}{e(X_i)\left(1 - e(X_i)\right)}\right] \tag{37}$$

for any policy class Π with finite VC dimension. Here, we also note that (14) implies that $V_{\max} \sim V\left(\pi^*\right)$ in our problem as the treatment effect gets small for large n. The following result shows that (37) is sharp up to constants.

Theorem 7. *Let $m(x)$, $e(x)$, and $\sigma(x)$ be functions over $\mathcal{X}_s$ satisfying the conditions discussed above, and let d be a positive integer. Then, there exists a class of functions Π over $\mathcal{X}_s$ with $\text{VC}\left(\Pi\right) = d$ (and a constant C) such that the minimax risk for policy learning over the data generating distribution (36) (with unknown $|\tau(x)| \leq C$) and the policy class Π is*

[19] For a precise argument, we need 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 (36) we have $\mathbb{E}\left[Y_i | X_i = x\right] = m(x)$ regardless of n. Thus, because both $e(x)$ and $m(x)$ are sufficiently differentiable, we can use standard results about series estimation to obtain $o_P(n^{-1/4})$-consistent estimators $\hat{e}(x)$ and $\hat{m}(x)$ for these quantities. Next, for the purpose of our policy learner, we simply set $\hat{\mu}_0(x) = \hat{\mu}_1(x) = \hat{m}(x)$; and because $\mathbb{E}\left[\tau^2(X_i)/\sqrt{n}\right] = \mathcal{O}(1/n)$, these regression adjustments in fact satisfy Assumption 1.

bounded from below as

$$\liminf_{n\to\infty}\left\{\sqrt{n}\inf_{\hat{\pi}_n}\left\{\sup_{|\tau(x)|\leq C}\left\{\mathbb{E}\left[R_n\left(\hat{\pi}_n\right)\right]\right\}\right\}\right\}\geq 0.33\sqrt{V_* d}. \tag{38}$$

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, 2017), 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 efficient policy evaluation for policy learning becomes apparent.

5 Discussion

In this paper, we showed how classical concepts from the literature on semiparametric efficiency can be used to develop performant algorithms for policy learning with strong asymptotic guarantees. Our regret bounds may prove to be particularly relevant in applications since, unlike existing bounds, they are sharp enough to distinguish between different a priori reasonable policy learning schemes (e.g., ones based on inverse-propensity weighting versus double machine learning), and thus provide methodological guidance to practitioners. We end our paper by discussing some potential extensions to our analysis.

First, following Manski (2004) and Hirano and Porter (2009), Kitagawa and Tetenov (2015), Stoye (2009), etc., our analysis is built on minimax regret bounds for learning a decision rule π from a pre-specified class Π that encodes constraints related to fairness, budget, functional form, etc. A limitation of this minimax approach is that it doesn't allow us to leverage further regularity properties of the optimal policy $\pi^* = \text{argmax}\{Q(\pi) : \pi \in \Pi\}$: For example, if Π consists of all k-sparse linear decision rules, but the optimal policy is actually k'-sparse for some $k' \ll k$, then our regret bounds will depend on k rather than k'.

Developing methods for policy learning that can adapt to the complexity of the optimal treatment allocation rule π^* would be of considerable interest. As one step in this direction, Mbakop and Tabord-Meehan (2016) extend the analysis of Kitagawa and Tetenov (2015) to the setting where an analyst wants to learn a policy π belonging to a sequence of nested policy classes Π_ℓ for $\ell = 1, 2, ...$, and consider the resulting problem of model selection (i.e., choosing the optimal index ℓ to use for empirical welfare maximization). Because our regret bounds hold uniformly over policy classes of different sizes satisfying our Assumption 2, we can pair our results about efficient policy learning with the model selection method of Mbakop and Tabord-Meehan (2016) to obtain improved regret bounds in their setting. Further work on adaptive policy learning would complement a long existing literature on adaptive minimax estimation and inference (e.g., Armstrong and Kolesár, 2016; Birgé and Massart, 2001; Cai and Low, 2005; Donoho and Johnstone, 1994; Efron, 1983; Lepskii, 1991).

Second, this paper has focused on policy learning in observational designs where the unconfoundedness assumption (2) holds. Although this setting is commonly studied, it is not the only possible setting where optimal policies may be estimated using observational data.

For example, in survival analysis, we may want to learn a treatment policy that maximizes expected quality-adjusted life years; and in order to do so, have access to an unconfounded treatment assignment mechanism but with incomplete outcomes due to administrative censoring (i.e., some subjects are lost to follow-up before death). Another example that arises frequently in econometrics is working with treatment effects that can only be identified via instruments.[20]

Because our analysis is framed in terms of uniform bounds for efficient policy evaluation, our results allow for fairly straight-forward extensions to more general settings. Given any problem where we know the efficient score for policy evaluation, we could learn $\hat{\pi}$ via a variant of (9) with weights $\widehat{\Gamma}_i$ obtained via an appropriate Neyman-orthogonal double-machine-learning construction as discussed by Chernozhukov et al. (2016). Our concentration bounds for the oracle policy learner (Theorems 3 and 6) would then hold verbatim, and we would only need to extend the argument from Lemma 4 that controls the uniform convergence of double machine learning.

Finally, our experience shows that results on semiparametrically efficient estimation are not just useful for statistical inference, but are also directly relevant to applied decision making problems. It will be interesting to see whether related insights will prove to be more broadly helpful for, e.g., sequential problems with contextual bandits, or non-discrete decision making problems involving, say, price setting or capacity allocation.

6 Proofs

6.1 Proof of Lemma 2

Our proof of this result follows the outline of Dudley's chaining argument, whereby we construct a sequence of approximating sets of increasing precision for $\widetilde{Q}_n(\pi)$ with $\pi \in \Pi_n$, and then use finite concentration inequalities to establish the behavior of $\widetilde{Q}_n(\pi)$ over this approximation set. The improvements in our results relative to existing bounds described in the body of the text come from a careful construction of approximating sets targeted to the problem efficient policy evaluation—for example, our use of chaining with respect to the random distance measure defined in (39)—and the use of sharp concentration inequalities.

Given these preliminaries, we start by defining the conditional 2-norm distance between two policies π_1, π_2 as (throughout this proof, we suppress the dependence of $\Gamma_i^{(n)}$ on n)

$$D_n^2(\pi_1, \pi_2) = \frac{1}{4}\sum_{i=1}^{n} \Gamma_i^2 \left(\pi_1(X_i) - \pi_2(X_i)\right)^2 \Big/ \sum_{i=1}^{n} \Gamma_i^2, \tag{39}$$

and let $N_{D_n}(\varepsilon, \Pi_n, \{X_i, \Gamma_i\})$ be the ε-covering number in this distance. To bound N_{D_n}, imagine creating another sample $\{X'_j\}_{j=1}^m$, with X'_j contained in the support of $\{X_i\}_{i=1}^n$, such that

$$\left| \left|\left\{j \in 1, ..., m : X'_j = X_i\right\}\right| - m\,\Gamma_i^2 \Big/ \sum_{i=1}^{n} \Gamma_i^2 \right| \leq 1.$$

[20] In order to learn optimal policies with treatment effect estimates specified via instruments, it would be helpful to make assumptions that let us identify the average welfare improvement of any policy over random assignment, rather than just a local average welfare improvement (Imbens and Angrist, 1994). For example, we could assume that treatment effects are conditionally uncorrelated with compliance, such that the conditional average treatment effect is identified as $\tau(x) = \text{Cov}\left[Y_i, Z_i \,\middle|\, X_i = x\right] / \text{Cov}\left[W_i, Z_i \,\middle|\, X_i = x\right]$, where Z_i is an instrument. If we cannot identify average welfare gains from policies, further conceptual developments may be necessary.

We immediately see that, for any two policies π_1 and π_2,

$$\frac{1}{m}\sum_{j=1}^{m} 1\left(\left\{\pi_1(X'_j) \neq \pi_2(X'_j)\right\}\right) = D_n^2\left(\pi_1,\, \pi_2\right) + \mathcal{O}\left(\frac{1}{m}\right).$$

Now, recall that N_H as used in our entropy integral (22) does not depend on sample size, so we can without reservations make m arbitrarily large, and conclude that

$$N_{D_n}\left(\varepsilon,\, \Pi_n,\, \{X_i,\, \Gamma_i\}\right) \leq N_H\left(\varepsilon^2,\, \Pi_n\right). \tag{40}$$

In other words, we have found that we can bound the D_n-entropy of Π_n with respect to its distribution-independent Hamming entropy.

Now, for every element $\pi \in \Pi_n$, define a set of approximations $A_j(\pi) : \mathcal{X} \to \{\pm 1\}$ for $j = 1,\, 2,\, ...$ with the property that $D_n(A_j(\pi),\, A_{j+1}(\pi)) \leq 2^{-j}$, and that the set $\Pi_n^j := \{A_j(\pi) : \pi \in \Pi_n\}$ of j-th order approximating policies has cardinality at most $N_{D_n}(2^{-j},\, \Pi_n,\, \{X_i,\, \Gamma_i\})$. Moreover, without loss of generality, we can construct these approximations such that there is no branching in the approximating sequences, i.e., $A_j(\pi) = A_j\left(A_{j+1}(\pi)\right)$ for all j and π. Finally, we use the notation $A_0(\pi)(x) = 0$. Then, for any index J, we clearly have

$$\pi(x) = \left(\pi(x) - A_J(\pi)(x)\right) + \sum_{j=1}^{J}\left(A_j(\pi)(x) - A_{j-1}(\pi)(x)\right), \tag{41}$$

and also note that, for any π and j,

$$\begin{aligned}\frac{1}{n}\,\mathrm{Var}\left[\sum_{i=1}^{n}\Gamma_i Z_i\left(A_j(\pi)(X_i) - A_{j+1}(\pi)(X_i)\right) \,\middle|\, \{X_i,\, \Gamma_i\}_{i=1}^n\right] \\ = 4\widehat{V} D_n^2\left(A_j(\pi)(X_i),\, A_{j+1}(\pi)(X_i)\right) \leq 2^{2-2j}\widehat{V},\end{aligned} \tag{42}$$

where $\widehat{V} = \sum_{i=1}^n \Gamma_i^2/n$.

Our goal is to use the series-based representation in (41) to obtain concentration bounds for our empirical process. To do so, it is helpful to consider terms on 3 different scales, set apart by

$$J(n) := \left\lfloor \log_2(n)\left(1 - \beta - \omega\right) \,/\, 2 \right\rfloor, \quad J_+(n) := \left\lfloor \log_2(n)\left(1 - \beta - \omega\right) \right\rfloor. \tag{43}$$

Given these thresholds, we can immediately use Jensen's inequality to check that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Gamma_i Z_i\left(\pi(X_i) - A_{J_+(n)}(\pi)(X_i)\right) &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Gamma_i^2\left(\pi(X_i) - A_{J_+(n)}(\pi)(X_i)\right)^2} \\ &= 2D_n\left(\pi(X_i),\, A_{J_+(n)}(\pi)(X_i)\right)\sqrt{\widehat{V}} \leq 2^{4-J_+(n)}\sqrt{\widehat{V}},\end{aligned}$$

and so, given that $\liminf J_+(n)/\log_2(n) > 1/2$ (because $\beta + \omega < 1/2$),

$$\lim_{n\to\infty}\sqrt{n}\,\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Gamma_i Z_i\left(\pi(X_i) - A_{J_+(n)}(\pi)(X_i)\right)\right] = 0,$$

meaning that we can safely ignore all terms $A_j(\pi)(X_i) - A_{j+1}(\pi)(X_i)$ with $j \geq J_+(n)$. The rest of the proof will show that the terms with $J(n) \leq j < J_+(n)$ are also asymptotically

negligible, while the low-order terms with $1 \leq j < J(n)$ determine the first-order behavior of $\mathcal{R}_n(\Pi_n)$.

Our arguments build on Bernstein's inequality, which guarantees that for any independent, mean-zero variables S_i with $|S_i| \leq M$, and any constant $t > 0$

$$\mathbb{P}\left[\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} S_i\right| \geq t\right] \leq 2\exp\left[\frac{-t^2}{2} \Big/ \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[S_i^2\right] + \frac{Mt}{3\sqrt{n}}\right)\right]. \tag{44}$$

In our problem, Bernstein's inequality means that conditionally on $\{X_i, \Gamma_i\}$ and writing $M_n = 2\sup\{|\Gamma_i| : 1 \leq i \leq n\}$, we have, for any choice of $t > 0$, $\pi \in \Pi_n$ and $j = 1, 2, ...,$

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i Z_i\left(A_j(\pi)(X_i) - A_{j+1}(\pi)(X_i)\right)\right| \geq t\,2^{2-j}\sqrt{\widehat{V}}\right] \qquad (45)\\
&\leq 2\exp\left[\frac{-t^2 4^{2-j}\widehat{V}}{2} \Big/ \left(\frac{4}{n}\sum_{i=1}^{n}\Gamma_i^2 1\left(\{A_j(\pi)(X_i) \neq A_{j+1}(\pi)(X_i)\}\right) + \frac{M_n t\,2^{2-j}\sqrt{\widehat{V}}}{3\sqrt{n}}\right)\right]\\
&= 2\exp\left[\frac{-t^2 4^{2-j}\widehat{V}}{2} \Big/ \left(16 D_n^2\left(A_j(\pi),\, A_{j+1}(\pi)\right)\widehat{V} + \frac{M_n t\,2^{2-j}\sqrt{\widehat{V}}}{3\sqrt{n}}\right)\right]\\
&\leq 2\exp\left[\frac{-t^2}{2}\left(1 + \frac{1}{12}\frac{M_n t\,2^j}{\sqrt{n\widehat{V}}}\right)^{-1}\right],
\end{aligned}$$

where on the last line we used the fact that $D_n^2\left(A_j(\pi),\, A_{j+1}(\pi)\right) \leq 4^{-j}$. Moreover, by the same argument

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i Z_i\left(A_j(\pi)(X_i) - \pi(X_i)\right)\right| \geq t\,2^{3-j}\sqrt{\widehat{V}}\right]\\
&\qquad\leq 2\exp\left[\frac{-t^2}{2}\left(1 + \frac{1}{24}\frac{M_n t\,2^j}{\sqrt{n\widehat{V}}}\right)^{-1}\right],
\end{aligned} \tag{46}$$

because $D_n^2\left(A_j(\pi),\, \pi\right) \leq 4^{1-j}$ for any policy π via the geometric series formula.

Given these preliminaries, we are now ready to verify that terms $A_j(\pi)(X_i) - A_{j+1}(\pi)(X_i)$ in (41) with $J(n) \leq j < J_+(n)$ are in fact negligible. To do so, we collapse all approximating policies with $J(n) \leq j < J_+(n)$, and directly compare $A_{J(n)}(\pi)$ to $A_{J_+(n)}(\pi)$. Because of our "no branching" construction, we know that $A_{J(n)}(\pi) = A_{J(n)}(A_{J_+(n)}(\pi))$ for all policies $\pi \in \Pi_n$, and so

$$\begin{aligned}
&\mathbb{P}\left[\sup\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i Z_i\left(A_{J(n)}(\pi)(X_i) - A_{J_+(n)}(X_i)\right)\right| : \pi \in \Pi_n\right\} \geq t\,2^{3-J(n)}\sqrt{\widehat{V}}\right]\\
&\quad= \mathbb{P}\left[\sup\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i Z_i\left(A_{J(n)}(\pi)(X_i) - \pi(X_i)\right)\right| : \pi \in \Pi_n^{J_+(n)}\right\} \geq t\,2^{3-J(n)}\sqrt{\widehat{V}}\right]\\
&\quad\leq 2\left|\Pi_n^{J_+(n)}\right|\exp\left[\frac{-t^2}{2}\left(1 + \frac{1}{24}\frac{M_n t\,2^{J(n)}}{\sqrt{n\widehat{V}}}\right)^{-1}\right],
\end{aligned}$$

where the last inequality is simply a union bound over (46). By Assumption, 2, we know that

$$\log\left|\Pi_n^{J_+(n)}\right| \leq \log N_{D_n}\left(2^{-J_+(n)},\, \Pi_n,\, \{X_i,\, \Gamma_i\}\right) \leq \log N_H\left(4^{-J_+(n)},\, \Pi_n\right) \leq Cn^{\beta}4^{\omega J_+(n)}.$$

Moreover, given our choice of $J_+(n)$ from (43), we get

$$n^{\beta}4^{\omega J_+(n)} \leq n^{\beta+2\omega(1-\beta-\omega)} = n^{\frac{1}{2}+\frac{(1-2\beta-2\omega)(2\omega-1)}{2}},$$

and note that $(1-2\beta-2\omega)(2\omega-1) < 0$ because $\omega,\, \beta+\omega < 1/2$ by Assumption 2. Thus, plugging $t^2 = 4^{J(n)}/(\widehat{V}\log(n))$ into the above bound we see that, for large values of n

$$\begin{aligned}
&\mathbb{P}\left[\sup\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i Z_i\left(A_{J(n)}(\pi)(X_i) - A_{J_+(n)}(X_i)\right)\right| : \pi \in \Pi_n\right\} \geq 8/\sqrt{\log(n)}\right] \\
&\quad\leq 2\left|\Pi_n^{J_+(n)}\right|\exp\left[\frac{-6t\sqrt{n\widehat{V}}}{2^{J(n)}M_n}\right] \\
&\quad\leq 2\exp\left[\sqrt{n}\left(Cn^{\frac{(1-2\beta-2\omega)(2\omega-1)}{2}} - \frac{6}{M_n\sqrt{\log(n)}}\right)\right].
\end{aligned}$$

Finally, noting that

$$M_n = \mathcal{O}_P\left(\sqrt{\log(n)}\right) \tag{47}$$

because Γ_i is sub-Gaussian and $\mathrm{Var}[\Gamma_i]$ is bounded from below (recall that M_n is the supremum of $|\Gamma_i|$), we conclude that $n^{(1-2\beta-2\omega)(2\omega-1)/2} \ll 6\,/\,M_n\sqrt{\log(n)}$ for large values of n, and so the right-hand side probability bound converges to 0. In other words, we have shown that

$$\sup\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i Z_i\left(A_{J(n)}(\pi)(X_i) - A_{J_+(n)}(X_i)\right)\right| : \pi \in \Pi_n\right\} \leq \frac{8}{\sqrt{\log(n)}}$$

with probability tending to 1 at a rate of $e^{-\Omega\left(\sqrt{n}/\log(n)\right)}$. Noting the speed of the convergence, it is also straight-forward to check that the expectation of the supremum is bounded at the same scale, and so term with $J(n) \leq j < J_+(n)$ in fact do not contribute to the Rademacher complexity, as claimed.

We are now finally ready to study the terms of (41) that matter, i.e., those with $j < J(n)$. To get started, for every $n,\, j \geq 1$ and a sequence $\delta_n > 0$, define the event

$$\begin{aligned}
\mathcal{E}_{j,\,n} &:= \left\{\sup_{\pi\in\Pi_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_i Z_i\left(A_j(\pi)(X_i) - A_{j+1}(\pi)(X_i)\right)\right| \geq 2^{2-j}t_{j,n}\sqrt{\widehat{V}}\right\} \\
t_{j,n} &:= 2\sqrt{\log\left(N_H\left(4^{-(j+1)},\, \Pi_n\right)\right) + \log\left(\frac{2j^2}{\delta_n}\right)}.
\end{aligned} \tag{48}$$

By (45), we immediately see that

$$\mathbb{P}\left[\mathcal{E}_{j,\,n}\right] \leq 2\left|\Pi_n^{j+1}\right|\exp\left[\frac{-t_{j,n}^2}{2}\left(1 + \frac{1}{12}\,\frac{M_n t_{j,n}\,2^j}{\sqrt{n\widehat{V}}}\right)^{-1}\right].$$

Moreover, using similar arguments to those made above (and in particular, leveraging Assumption 2) we see that, for all $j \leq J(n) - 1$,

$$\frac{\sqrt{\log\left(N_H\left(4^{-(j+1)},\ \Pi_n\right)\right)}2^j}{\sqrt{n}} \leq \frac{\sqrt{\log\left(N_H\left(4^{-J(n)},\ \Pi_n\right)\right)}2^{J(n)}}{\sqrt{n}}$$
$$\leq \frac{\left(Cn^{\beta/2}2^{\omega J(n)}\right)2^{J(n)}}{\sqrt{n}} \leq n^{\frac{\beta+(1+\omega)(1-\beta-\omega)-1}{2}} = n^{\frac{-\omega(\beta+\omega)}{2}}.$$

Recalling (47) and assuming that δ_n^{-1} grows at most polynomially in n, this implies that there is an index N such that for all $n \geq N$ and all $j < J(n)$,

$$\mathbb{P}\left[\mathcal{E}_{j,\,n}\right] \leq 2\left|\Pi_n^{j+1}\right| \exp\left[\frac{-t_{j,n}^2}{4}\right] = \frac{\delta_n}{j^2}\left|\Pi_n^{j+1}\right| \Big/ N_H\left(4^{-(j+1)},\ \Pi_n\right) \leq \frac{\delta_n}{j^2}.$$

Thus, we find that, for large enough n, $\mathcal{E}_{j,\,n}$ does not happen for any $1 \leq j < J(n)$ with probability at least $1 - \delta_n \sum_{j=1}^{\infty} j^{-2}$.

Moreover, on the event than none of these $\mathcal{E}_{j,\,n}$ happen,

$$\sqrt{n}\sup_{\pi\in\Pi_n}\left|\frac{1}{n}\sum_{i=1}^{n}\Gamma_i Z_i\sum_{j=1}^{J(n)}\left(A_j(\pi) - A_{j-1}(\pi)\right)(X_i)\right| \tag{49}$$
$$\leq \sqrt{\widehat{V}}\sum_{j=1}^{J(n)} 2^{3-j}\sqrt{\log\left(N_H\left(4^{-j},\ \Pi_n\right)\right) + \log(j^2/\delta_n)}$$
$$\leq 8\sqrt{\widehat{V}}\int_0^1 \sqrt{\log\left(N_H\left(\varepsilon^2,\ \Pi_n\right)\right) + 2\log\left(1 + \log_2\left(\varepsilon^{-1}\right)\right) + \log(\delta_n^{-1})}\ d\varepsilon$$
$$\leq 8\sqrt{\widehat{V}}\left(\kappa\left(\Pi_n\right) + \int_0^1 \sqrt{2\log\left(1 + \log_2\left(\varepsilon^{-1}\right)\right)}\ d\varepsilon + \sqrt{\log(\delta_n^{-1})}\right).$$

Applying this bound separately for the sequences $\delta_n = \max\left\{2^{-k},\ 1/n\right\}$ for $k = 1,\ 2,\ ...$, we can turn the above into a bound on the expectation:

$$\sqrt{n}\,\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\frac{1}{n}\sum_{i=1}^{n}\Gamma_i Z_i\sum_{j=1}^{J(n)}\left(A_j(\pi) - A_{j-1}(\pi)\right)(X_i)\right|\right]$$
$$\leq 8\mathbb{E}\left[\sqrt{\widehat{V}}\right]\left(\kappa\left(\Pi_n\right) + \int_0^1 \sqrt{2\log\left(1 + \log_2\left(\varepsilon^{-1}\right)\right)}\ d\varepsilon + \sum_{k=1}^{\infty}\frac{\sqrt{k\log(2)}}{2^k}\right) + \mathcal{O}\left(\sqrt{\frac{\log(n)}{n}}\right),$$

where the last term is a crude bound (obtained via (47)) on the expectation of our statistic of interest on the event that all Bernstein bounds fail, occurring with probability at most $1/n$. Finally, recalling our earlier conclusion that higher order terms in the expansion (41) do not asymptotically affect the Rademacher complexity, we conclude that

$$\mathcal{R}_n(\Pi_n) \leq 8\sqrt{\frac{\mathbb{E}\left[\Gamma^2\right]}{n}}\left(\kappa\left(\Pi_n\right) + \int_0^1 \sqrt{2\log\left(1 + \log_2\left(\varepsilon^{-1}\right)\right)}\ d\varepsilon + \sum_{k=1}^{\infty}\frac{\sqrt{k\log(2)}}{2^k}\right) + \mathcal{O}\left(\sqrt{\frac{\log(n)}{n}}\right),$$

noting that $\mathbb{E}[\sqrt{\widehat{V}}] \leq \sqrt{\mathbb{E}\left[\Gamma^2\right]}$ by concavity of the square-root function.

6.2 Proof of Theorem 3

First, as argued by, e.g., Bartlett and Mendelson (2002) in the proof of their Theorem 8,

$$\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\widetilde{Q}_n(\pi)-Q_n(\pi)\right|\right]\leq 2\mathbb{E}\left[\mathcal{R}_n(\Pi_n)\right]. \tag{50}$$

Thus, to make use of Lemma 2, it suffices to bound $\sup_{\pi\in\Pi_n}\left|\widetilde{Q}_n(\pi)-Q_n(\pi)\right|$ in terms of its expectation. Now, recall that $\widetilde{Q}_n(\pi)=n^{-1}\sum\Gamma_i^{(n)}\pi(X_i)$, and that the $\Gamma_i^{(n)}$ are uniformly sub-Gaussian. Now, because the $\Gamma_i^{(n)}$ are not bounded, it is convenient to define truncated statistics

$$\widetilde{Q}_n^{(-)}(\pi)=\frac{1}{n}\sum_{i=1}^{n}\Gamma_i^{(n-)}\pi(X_i),\quad \Gamma_i^{(n-)}=\Gamma_i^{(n)}\,\mathbf{1}\left(\left\{\left|\Gamma_i^{(n)}\right|\leq\log(n)\right\}\right).$$

Here, we of course have that $|\Gamma_i^{(n-)}|\leq\log(n)$, and so we can apply Talagrand's inequality as described in Bousquet (2002) to these truncated statistics. We see that, for any $\delta>0$, with probability at least $1-\delta$,

$$\begin{aligned}\sup_{\pi\in\Pi_n}\left|\widetilde{Q}_n^{(-)}(\pi)-Q_n^{(-)}(\pi)\right|&\leq\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\widetilde{Q}_n^{(-)}(\pi)-Q_n^{(-)}(\pi)\right|\right]\\&+\sqrt{2\frac{\log(\delta)}{n}\left(\mathbb{E}\left[\left(\Gamma_i^{(n)}\right)^2\right]+2\log(n)\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\widetilde{Q}_n^{(-)}(\pi)-Q_n^{(-)}(\pi)\right|\right]\right)}+\frac{\log(n)\log(\delta)}{3n},\end{aligned}$$

where we used the short-hand $Q_n^{(-)}(\pi)=\mathbb{E}[\widetilde{Q}_n^{(-)}(\pi)]$. Moreover, because the $\Gamma_i^{(n)}$ are uniformly sub-Gaussian, we can immediately verify that

$$\mathbb{E}\left[\left|\sup_{\pi\in\Pi_n}\left|\widetilde{Q}_n^{(-)}(\pi)-Q_n^{(-)}(\pi)\right|-\sup_{\pi\in\Pi_n}\left|\widetilde{Q}_n(\pi)-Q_n(\pi)\right|\right|\right]$$

decays exponentially fast in n. Using (50) and noting that, by Lemma 2, $\mathbb{E}\left[\mathcal{R}_n(\Pi_n)\right]$ decays as $\mathcal{O}(1/\sqrt{n})$, we conclude that with probability at least $1-\delta$,

$$\begin{aligned}\sup_{\pi\in\Pi_n}\left|\widetilde{Q}_n(\pi)-Q_n(\pi)\right|&\leq 2\mathbb{E}\left[\mathcal{R}_n(\Pi_n)\right]\\&+\sqrt{2\mathbb{E}\left[\left(\Gamma_i^{(n)}\right)^2\right]\log(\delta)\,/\,n}+\mathcal{O}\left(\frac{\log(n)\mathcal{R}_n(\Pi_n)}{\sqrt{n}}\right),\end{aligned} \tag{51}$$

thus establishing the first part of the theorem statement. Meanwhile, to prove the second part, we simply note that $\widetilde{Q}_n(\tilde{\pi}_n)\geq\widetilde{Q}_n(\pi_n^*)$ by construction, and then apply (51) at both $\tilde{\pi}_n$ and π_n^* (where π_n^* denotes the regret-minimizing policy in the n-th problem of our sequence).

6.3 Proof of Lemma 4

To streamline notation, we omit (n)-superscripts on $\hat{\mu}(\cdot)$, $\hat{e}(\cdot)$, $\widehat{Q}$, etc., throughout this proof. For any fixed policy π, we begin by expanding out the difference of interest. Write

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

and define $\widehat{Q}_{-1}(\pi)$ and $\widetilde{Q}_{\pm 1}(\pi)$ analogously, such that $\widehat{Q}(\pi) = \widehat{Q}_{+1}(\pi) - \widehat{Q}_{-1}(\pi)$, etc. Then,

$$\begin{aligned}
\widehat{Q}_{+1}(\pi) - \widetilde{Q}_{+1}(\pi) &= \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\Bigg(\hat{\mu}_{+1}^{(-k(i))}(X_i) - \mu_{+1}(X_i) \\
&\qquad + \mathbf{1}\left(\{W_i = +1\}\right)\left(\frac{Y_i - \hat{\mu}_{+1}^{(-k(i))}(X_i)}{\hat{e}_{+1}^{(-k(i))}(X_i)} - \frac{Y_i - \mu_{+1}(X_i)}{e_{+1}(X_i)}\right)\Bigg) \\
&= \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\left(\hat{\mu}_{+1}^{(-k(i))}(X_i) - \mu_{+1}(X_i)\right)\left(1 - \frac{\mathbf{1}\left(\{W_i = 1\}\right)}{e_{+1}(X_i)}\right) \\
&\qquad + \frac{1}{n}\sum_{\{i:W_i=1\}} \pi(X_i)\left(Y_i - \mu_{+1}(X_i)\right)\left(\frac{1}{\hat{e}_{+1}^{(-k(i))}(X_i)} - \frac{1}{e_{+1}(X_i)}\right) \\
&\qquad + \frac{1}{n}\sum_{\{i:W_i=1\}} \pi(X_i)\left(\mu_{+1}(X_i) - \hat{\mu}_{+1}^{(-k(i))}(X_i)\right)\left(\frac{1}{\hat{e}_{+1}^{(-k(i))}(X_i)} - \frac{1}{e_{+1}(X_i)}\right).
\end{aligned}$$

Denote these three summands by $A_{+1}(\pi)$, $B_{+1}(\pi)$, $C_{+1}(\pi)$. We will be to bound all 3 summands separately.

To bound the first term, it is helpful separate out the contributions of the K different folds:

$$A_{+1}^{(k)}(\pi) = \frac{1}{n}\sum_{\{i:k(i)=k\}} \pi(X_i)\left(\hat{\mu}_{+1}^{(-k)}(X_i) - \mu_{+1}(X_i)\right)\left(1 - \frac{\mathbf{1}\left(\{W_i = 1\}\right)}{e_{+1}(X_i)}\right). \tag{52}$$

Now, because $\hat{\mu}_{+1}^{(-k)}(\cdot)$ was only computed using data from the $K-1$ folds, we can condition on the value of this function estimate to make the individual terms in the above sum independent. By Assumption 1, we know that

$$\sup_{x\in\mathcal{X}} \left|\left(\hat{\mu}_{+1}^{(-k)}(x) - \mu_{+1}(x)\right)\right| \leq \eta$$

with probability tending to 1, and so, by overlap, the individual summands in (52) are bounded by 1 (and thus uniformly sub-Gaussian) with probability tending to 1. Then, writing

$$V_n(k) = \mathbb{E}\left[\left(\hat{\mu}_{+1}^{(-k)}(X) - \mu_{+1}(X)\right)^2\left(1 - \frac{\mathbf{1}\left(\{W_i = 1\}\right)}{e_{+1}(X_i)}\right)^2 \,\middle|\, \hat{\mu}_{+1}^{(-k)}(\cdot)\right]$$

for the variance of $A_{+1}^{(k)}(\pi)$ conditionally on the regression model $\hat{\mu}_{+1}^{(-k)}(\cdot)$ fit on the other $K-1$ folds, we can apply Theorem 3 to establish that

$$\frac{n}{n_k}\sup_{\pi\in\Pi}\left|A_{+1}^{(k)}(\pi)\right| \,\Bigg|\, \hat{\mu}_{+1}^{(-k)}(\cdot) = \mathcal{O}_P\left(\kappa\left(\Pi_n\right)\sqrt{\frac{V_n(k)}{n_k}}\right) + \mathcal{O}\left(\frac{\log\left(n_k\right)}{n_k}\right), \tag{53}$$

where $n_k = |\{i : k(i) = k\}|$ denotes the number of observations in the k-th fold.

Next, recalling that constructed our double machine learning estimator using a finite number of evenly-sized folds, $n_k/n \to 1/K$, we can use overlap (for the first inequality below) and our risk bounds in Assumption 1 (for the second inequality) to check that

$$V_n(k) \leq \frac{1}{\eta^2}\,\mathbb{E}\left[\left(\hat{\mu}_{+1}^{(-k)}(X) - \mu_{+1}(X)\right)^2 \,\middle|\, \hat{\mu}_{+1}^{(-k)}(\cdot)\right] = \mathcal{O}_P\left(a\left(\frac{K-1}{K}\,n\right)\,n^{-\zeta_\mu}\right). \tag{54}$$

Then, applying (53) separately to all K folds and using Markov's inequality, we find that

$$\begin{aligned} A_{+1}(\pi) &= \mathcal{O}_P\left(\kappa\left(\Pi_n\right)\sqrt{\frac{a((1-K^{-1})n)}{n^{1+\zeta_\mu}}}\right), \\ B_{+1}(\pi) &= \mathcal{O}_P\left(\kappa\left(\Pi_n\right)\sqrt{\frac{a((1-K^{-1})n)}{n^{1+\zeta_e}}}\right), \end{aligned} \tag{55}$$

noting that the argument used to bound $B_{+1}(\pi)$ is analogous to the one used for $A_{+1}(\pi)$.

It now remains to bound the final term, $C_{+1}(\pi)$. Here, we can use the Cauchy-Schwarz inequality to verify that

$$\begin{aligned} C_{+1}(\pi) &= \frac{1}{n}\sum_{\{i:W_i=1\}} \pi(X_i)\left(\mu_{+1}(X_i) - \hat{\mu}_{+1}^{-k(i)}(X_i)\right)\left(\frac{1}{\hat{e}_{+1}^{-k(i)}(X_i)} - \frac{1}{e_{+1}(X_i)}\right) \\ &\leq \sqrt{\frac{1}{n}\sum_{\{i:W_i=1\}}\left(\mu_{+1}(X_i) - \hat{\mu}_{+1}^{-k(i)}(X_i)\right)^2}\sqrt{\frac{1}{n}\sum_{\{i:W_i=1\}}\left(1\,/\,\hat{e}_{+1}^{-k(i)}(X_i) - 1\,/\,e_{+1}(X_i)\right)^2}. \end{aligned}$$

Then, applying Cauchy-Schwarz again to the above product, we see that

$$\begin{aligned} \mathbb{E}\left[\frac{n\,C_{+1}(\pi)}{|\{i:W_i=1\}|}\right] &\leq \sqrt{\mathbb{E}\left[\left(\hat{\mu}_{+1}^{-k(i)}(X) - \mu_{+1}(X)\right)^2\right]\mathbb{E}\left[\left(1/\hat{e}_{+1}^{-k(i)}(X) - 1/e_{+1}(X)\right)^2\right]} \\ &\leq a\left(\left\lfloor\frac{K-1}{K}\,n\right\rfloor\right)\Big/\sqrt{\left\lfloor\frac{K-1}{K}\,n\right\rfloor}, \end{aligned}$$

The desired conclusion now follows from Markov's inequality, along with an application of the same argument to $\widehat{Q}_{-1}(\pi)$.

6.4 Proof of Theorem 6

Throughout this proof, we suppress sub- and superscripts indexing dependence on n. To establish this result, we follow the strategy in the proof of Theorem 3. We apply Talagrand's inequality to get the following analogue to (51),

$$\sup\left\{\left|\widetilde{Q}\left(\pi\right) - Q\left(\pi\right)\right| : \pi \in \Pi_\lambda\right\} \leq 2\mathbb{E}\left[\mathcal{R}\left(\Pi_\lambda\right)\right] + \sqrt{\frac{2V_\lambda\log\left(\delta\right)}{n}} + \mathcal{O}\left(\frac{\log(n)}{n}\right), \tag{56}$$

and so our task again reduces to bounding the Rademacher complexity $\mathbb{E}\left[\mathcal{R}\left(\Pi_\lambda\right)\right]$. At this point, however, it is convenient to slightly alter our definition of Rademacher complexity, and set

$$\mathcal{R}\left(\Pi_\lambda\right) := \sup\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i\left(\Gamma_i\pi(X_i) - Q\left(\pi^*\right)\right)\right\}, \tag{57}$$

where the Z_i are independent Rademacher variables and the Γ_i are defined as before. Here, the addition of a constant offset by no means alters the argument behind (56); formally, we would get to this notion of Rademacher complexity by trying to establish concentration of $|\widetilde{Q}\left(\pi\right) - Q(\pi^*) - \left(Q\left(\pi\right) - Q(\pi^*)\right)|$. However, this offset term will let us leverage the fact that $\pi \in \Pi_\lambda$ to get better bounds.

Now, we start following the proof of Lemma 2 exactly up to (49), implying that

$$\mathcal{R}\left(\Pi_\lambda\right) = \left(1 + o_P(1)\right) \mathbb{E}\left[\sup_{\pi \in \Pi_\lambda} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i \left(\Gamma_i A_{J(n)}(\pi) - Q\left(\pi^*\right)\right)\right|\right], \tag{58}$$

with $J(n) = \lfloor \log_2(n)(3/2 - \beta)/4 \rfloor$ (this follows from (43) by setting $\omega = (1/2 - \beta)/2$, which satisfies Assumption 2 whenever Π is a VC-class and $\beta < 1/2$). Now, in the previous proof, we continued by writing the right-hand side of (58) as a chained sum in (49); here, in contrast, we need to use a more careful partial chaining argument instead.

As a preliminary step to doing so, define

$$J_0 := \max\left\{1, \lfloor \log_4(V_{\max}/V_\lambda) \rfloor\right\},$$

and note that we can define a new approximating function $A^\lambda_{J_0}(\cdot)$ with the following properties:

$$D\left(A^\lambda_{J_0}(\pi), A_{J_0+1}(\pi)\right) \leq 2^{-J_0} \text{ for all } \pi \in \Pi_\lambda, \quad A^\lambda_{J_0}(\pi) \in \Pi_\lambda \text{ for all } \pi \in \Pi_\lambda, \text{ and}$$
$$\left|\left\{A^\lambda_{J_0}(\pi) : \pi \in \Pi_\lambda\right\}\right| \leq \left|\left\{A_{J_0+1}(\pi) : \pi \in \Pi_\lambda\right\}\right| \leq N_H\left(4^{-J_0+1}, \Pi\right).$$

In order to build such an approximation, we can check that $D(\pi, A_{J_0+1}(\pi)) \leq 2^{-J_0}$ for all $\pi \in \Pi$ by construction. Thus, for every element $\pi' \in \{A_{J_0+1}(\pi) : \pi \in \Pi_\lambda\}$ we know that there must exist and element $\pi \in \Pi_\lambda$ for which $D(\pi, \pi') \leq 2^{-J_0}$. We can then define the approximating function $A^\lambda_{J_0}(\cdot)$ by mapping each unique element of $\{A_{J_0+1}(\pi) : \pi \in \Pi_\lambda\}$ to a policy in Π_λ using this relationship. We are now ready to proceed with our partial chaining argument, and write

$$\begin{aligned}\sup_{\pi \in \Pi_\lambda} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i \left(\Gamma_i A_{J(n)}(\pi) - Q\left(\pi^*\right)\right)\right| &\leq \sup_{\pi \in \Pi_\lambda} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i \left(\Gamma_i A^\lambda_{J_0}(\pi) - Q\left(\pi^*\right)\right)\right| \\ &+ \sup_{\pi \in \Pi_\lambda} \left| \frac{1}{n} \sum_{j=J_0+1}^{J(n)} \sum_{i=1}^{n} Z_i \Gamma_i \left(A^\lambda_j(\pi) - A^\lambda_{j-1}(\pi)\right)\right|,\end{aligned} \tag{59}$$

where we used the notational shorthand $A^\lambda_{J_0}(\pi) := A^\lambda_{J_0}(\pi)$ and $A^\lambda_j(\pi) := A_j(\pi)$ for $j \geq J_0+1$. Below, we bound both terms in (59) separately.

We start with the first term, and note that for any policy $\pi \in \Pi_\lambda$

$$\mathrm{Var}\left[Z_i \left(\Gamma_i \pi(X_i) - Q\left(\pi^*\right)\right) \mid \{X_i, \Gamma_i\}\right] = \frac{1}{n} \sum_{i=1}^{n} \left(\Gamma_i \pi(X_i) - Q\left(\pi^*\right)\right)^2.$$

Moreover, we can check that

$$\begin{aligned}\mathbb{E}\left[\left(\Gamma_i \pi(X_i) - Q\left(\pi^*\right)\right)^2\right] &= \mathrm{Var}\left[\Gamma_i \pi(X_i)\right] + \left(\mathbb{E}\left[\Gamma_i \pi(X_i)\right] - Q\left(\pi^*\right)\right)^2 \\ &= V(\pi) + R(\pi)^2 = V_* + 2Q\left(\pi^*\right) R(\pi) \leq V_\lambda,\end{aligned}$$

where all above above algebraic manipulations follow immediately from the definitions of the involved quantities (e.g., recall that $\widetilde{Q}(\pi) = \sum_i \pi(X_i)\Gamma_i/n$). Given these preliminaries,

a straight-forward application of Bernstein's inequality (44) tells us that, for any $\delta > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left[\frac{1}{\sqrt{nV_\lambda}} \sup_{\{\pi\in\Pi_\lambda\}} \left\{\left|\sum_{i=1}^n Z_i \left(\Gamma_i A^\lambda_{J_0}(\pi)(X_i) - Q\left(\pi^*\right)\right)\right|\right\} \geq 2\sqrt{\log\left(N_H\left(4^{-(J_0+1)}, \Pi\right)\right) + \log\left(\frac{\delta}{2}\right)}\right] \leq \delta.$$

Furthermore, continuing the same reasoning as before, we see that

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{\sqrt{nV_\lambda}} \mathbb{E}\left[\sup_{\{\pi\in\Pi_\lambda\}} \left\{\left|\sum_{i=1}^n Z_i \left(\Gamma_i A^\lambda_{J_0}(\pi)(X_i) - Q\left(\pi^*\right)\right)\right|\right\}\right] \\ \leq 2\sqrt{\log\left(N_H\left(4^{-(J_0+1)}, \Pi\right)\right)} + C, \\ \leq 2\sqrt{\alpha(J_0+1)\log(4)} + C,\end{aligned} \tag{60}$$

for some universal constant C, where on the last line we used our assumption (34).

Meanwhile, as to the second term, we can exactly follow the argument in Lemma 2 to verify that, for large enough n,

$$\begin{aligned}&\mathbb{E}\left[\sup_{\pi\in\Pi_\lambda}\left|\frac{1}{n}\sum_{j=J_0+1}^{J(n)}\sum_{i=1}^n Z_i\Gamma_i\left(A^\lambda_j(\pi) - A^\lambda_{j-1}(\pi)\right)\right|\right] \\ &\quad\leq 8\sqrt{V_{\max}}\int_0^{2^{1-J_0}}\left(\sqrt{\log\left(N_H\left(4\varepsilon^2, \Pi\right)\right)} + \sqrt{2\log\left(1+\log_2\left(\varepsilon^{-1}\right)\right)} + C\right)\,d\varepsilon \\ &\quad\leq 8\sqrt{V_{\max}}\left(\int_0^{2^{1-J_0}}\sqrt{2\alpha\log\left(\varepsilon^{-1}\right)} + \sqrt{2\log\left(1+\log_2\left(\varepsilon^{-1}\right)\right)}\,d\varepsilon + 2^{1-J_0}C\right),\end{aligned}$$

for some constant C; to establish the above inequality, we also used (34). We can then verify by calculus that the above expression can be bounded by

$$C\sqrt{\max\left\{\alpha, 1\right\}V_{\max}4^{1-J_0}\log\left(2^{1-J_0}\right)}$$

for some (new) constant C. To show this, it is helpful to note that

$$\int_0^t \sqrt{\log(1/\varepsilon)}\,d\varepsilon \leq 2t\sqrt{\log\left(1/t\right)} \text{ for any } 0 < t < 1/2.$$

Pulling everything together, we have found that

$$\begin{aligned}\sqrt{n}\,\mathbb{E}\left[\sup_{\pi\in\Pi_\lambda}\left|\frac{1}{n}\sum_{i=1}^n Z_i\left(\Gamma_i\pi(X_i) - Q\left(\pi^*\right)\right)\right|\right] &\leq C\sqrt{\max\left\{\alpha, 1\right\}\left(V_\lambda J_0 + V_{\max} J_0\, 4^{-J_0}\right)} \\ &\leq C\sqrt{5\max\left\{\alpha, 1\right\}V_\lambda \max\left\{1, \log_4(V_{\max}/V_\lambda)\right\}},\end{aligned}$$

thus establishing the desired result.

6.5 Proof of Theorem 7

We start by defining a specific choice of Π for which this bound holds. Let $0 = a_0 < a_1 < \ldots < a_d = 1$ and write $\mathcal{A}_j = \{x : a_{j-1} < x_1 < a_j\}$, $j = 1, ..., d$ for the induced partition of $\mathcal{X}_s$ along the first feature (we also assign points with $x_1 = 0$ to $\mathcal{A}_1$), such that

$$\mathbb{E}\left[1\left(\{X \in \mathcal{A}_j\}\right) \frac{\sigma^2(X)}{e(X)(1-e(X))}\right] = \frac{V_*}{d} \quad \text{for} \quad j = 1, ..., d. \tag{61}$$

Given this setup, we consider the policy class Π defined as the set of all 2^d policies that are piecewise constant over the sets $\mathcal{A}_j$ (i.e., each policy $\pi \in \Pi$ maps sets $\mathcal{A}_j$ entirely to either -1 or $+1$). Note that the VC-dimension of this class is trivially $\text{VC}(\Pi) = d$, because an arbitrary function class over a support of size d can shatter exactly d distinct points.

Now, to lower-bound the minimax risk for policy learning over all bounded treatment effect functions $\tau(\cdot)$, it is sufficient to bound minimax risk over a smaller class of policies T, as minimax risk increases with the complexity of the class T. Noting this fact, we restrict our analysis to treatment functions T such that

$$\tau(x) = \frac{\sigma^2(x)\, c_j}{e(x)(1-e(x))} \Big/ \mathbb{E}\left[\frac{\sigma^2(x)\, 1\left(\{X \in \mathcal{A}_j\}\right)}{e(X)(1-e(X))}\right]$$

for all $x \in \mathcal{A}_j$, where $c_j \in \mathbb{R}$ is an unknown coefficient for each $j = 1, ..., d$. If we knew the values of c_j for $j = 1, 2, ..., d$, the optimal policy $\pi^* \in \Pi$ would be treat only those j-groups with a positive c_j, i.e., $\pi^*(x) = \text{sign}(c_j)$ for all $x \in \mathcal{A}_j$.

Now, following the argument of Hirano and Porter (2009) (we omit details for brevity), the minimax policy learner is of the form $\hat{\pi}^*(x) = \text{sign}(\hat{c}_j^*)$ for all $x \in \mathcal{A}_j$, where $\hat{c}_j^*$ is an efficient estimator for c_j. Moreover, in this example, we can use (61) to readily verify that the semiparametric efficient variance for estimating c_j is V_*/d. Thus, the efficient estimator $\hat{c}_j^*$ will incorrectly estimate the sign of c_j with probability tending to $\Phi(-c_j\sqrt{d/V_*})$, where $\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function. (Recall that, in our sampling model (36), the signal also decays as $1/\sqrt{n}$.)

By construction, we suffer an expected utility loss of $2\,|c_j|$ from failing to accurately estimate the sign of c_j. Thus, by the above argument, given fixed values of c_j, the efficient policy learner will suffer an asymptotic regret

$$\lim_{n\to\infty} \sqrt{n}\mathbb{E}\left[R_n\right] = \sum_{j=1}^{d} 2\,|c_j|\,\Phi\left(-\,|c_j|\,\sqrt{d/V_*}\right),$$

assuming that an efficient estimator $\hat{c}_j^*$ in fact exists (and we know that one does, following the discussion in Section 4). Setting $|c_j| = 0.75\sqrt{V_*/d}$, this limit becomes

$$\lim_{n\to\infty} \sqrt{n}\mathbb{E}\left[R_n\right] = 1.5\Phi(-0.75)\sqrt{d\,V_*},$$

which, noting that $1.5\Phi(-0.75) \geq 0.33$, concludes the proof.

References

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1638–1646, 2014.

T. B. Armstrong and M. Kolesár. Optimal inference in a class of regression models. 2016.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016a.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *arXiv preprint arXiv:1610.01271*, 2016b.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

H. Bastani and M. Bayati. Online decision-making with high-dimensional covariates. 2015.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138. ACM, 2009.

D. Bhattacharya and P. Dupas. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196, 2012.

P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1998.

L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

L. Bottou, J. Peters, J. Q. Candela, D. X. Charles, M. Chickering, E. Portugaly, D. Ray, P. Y. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

T. T. Cai and M. G. Low. On adaptive estimation of linear functionals. *The Annals of Statistics*, 33(5):2311–2343, 2005.

G. Chamberlain. Bayesian aspects of treatment choice. In *The Oxford Handbook of Bayesian Econometrics*. 2011.

G. Chen, D. Zeng, and M. R. Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, (just-accepted), 2016.

L.-Y. Chen and S. Lee. Best subset binary prediction. *arXiv preprint arXiv:1610.02738*, 2016.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pages 442–450, 2010.

R. H. Dehejia. Program evaluation as a decision problem. *Journal of Econometrics*, 125(1): 141–173, 2005.

D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011.

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331, 1983.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

E. Greenshtein et al. Best subset selection, persistence in high-dimensional statistical learning and optimization under l1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

T. Grubinger, A. Zeileis, and K.-P. Pfeiffer. evtree: Evolutionary learning of globally optimal classification and regression trees in r. 61(1), 2014.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

K. Hirano and J. R. Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77 (5):1683–1701, 2009.

K. Hirano and J. R. Porter. Panel asymptotics and statistical decision theory. *The Japanese Economic Review*, 67(1):33–49, 2016.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

N. Kallus. Recursive partitioning for personalization using observational data. pages 1789–1798, 2017.

M. Kasy. Partial identification, distributional preferences, and the welfare ranking of policies. *Review of Economics and Statistics*, 98(1):111–131, 2016.

T. Kitagawa and A. Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 2015.

T. Kitagawa, A. Tetenov, et al. Equality-minded treatment choice. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 2017.

J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. Technical report, National Bureau of Economic Research, 2017.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

L. M. Le Cam. *Asymptotic Methods in Statistical Theory*. Springer-Verlag New York, Inc., 1986.

O. Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

A. Luedtke and A. Chambaz. Faster rates for policy learning. *arXiv preprint arXiv:1704.06431*, 2017.

C. F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72 (4):1221–1246, 2004.

C. F. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2009.

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

E. Mbakop and M. Tabord-Meehan. Model selection for treatment choice: Penalized welfare maximization. *arXiv preprint arXiv:1609.03167*, 2016.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 62(6):1349–1382, 1994.

J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. 41(2):693–721, 2013.

M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180, 2011.

A. Rakhlin and K. Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. 2016.

J. Robins, L. Li, R. Mukherjee, E. Tchetgen, and A. van der Vaart. Minimax estimation of a functional on a structured high dimensional model. *Annals of Statistics, forthcoming*, 2017.

J. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(1):122–129, 1995.

J. Robins, A. Rotnitzky, and L. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(1):106–121, 1995.

J. Robins, L. Li, E. Tchetgen, and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

L. J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46(253):55–67, 1951.

A. Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

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

J. Stoye. Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156, 2012.

X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, 10:141–158, 2009.

A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

A. Tetenov. Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 267–288, 1996.

M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media, 2011.

V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

S. Wager, W. Du, J. Taylor, and R. J. Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45): 12673–12678, 2016.

A. Wald. *Statistical Decision Functions*. Wiley, 1950.

B. Zhang, A. A. Tsiatis, M. Davidian, M. Zhang, and E. Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

X. Zhou, N. Mayer-Hamblett, U. Khan, and M. R. Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.