Efficient Policy Learning from Surrogate-Loss Classification Reductions

Andrew Bennett [1] **Nathan Kallus** [1]

Abstract

Recent work on policy learning from observational data has highlighted the importance of efficient policy evaluation and has proposed reductions to weighted (cost-sensitive) classification. However, efficient policy evaluation need not yield efficient estimation of optimal policy parameters. We consider the estimation problem given by a weighted surrogate-loss classification reduction of policy learning with any score function — either direct, inverse-propensity weighted, or doubly robust — and show that, under a correct specification assumption, the weighted classification formulation need not be efficient for policy parameters. We draw a contrast to actual (possibly weighted) binary classification, where correct specification implies a parametric model, while for policy learning it only implies a semiparametric model, and we show that efficiency in optimal parameter estimation implies optimal regret. In light of this, we instead propose an estimation approach based on the generalized method of moments, which is efficient for the policy parameters. We propose a particular method based on recent developments on solving moment problems using neural networks and demonstrate the efficiency and regret benefits of this method empirically.

1. Introduction

Policy learning from observational data is an important but challenging problem because it requires reasoning about the effects of interventions not observed in the data. For example, if we wish to learn an improved policy for medical treatment assignment based on observational data from electronic health records, we must take care to consider potential confounding: since healthier patients who were already predisposed to positive outcomes were likely to have historically been assigned less invasive treatments, naïve approaches may incorrectly infer that a policy of always assigning less invasive treatments will obtain better outcomes.

Various recent work has tackled this problem, known as policy learning from observational (or, off-policy) data, by optimizing causally-grounded estimates of policy value such as inverse-propensity weighting (IPW), doubly robust (DR) estimates, or similar (Beygelzimer & Langford, 2009; Jiang et al., 2019; Kallus, 2017; 2018; Kallus & Zhou, 2018; Kitagawa & Tetenov, 2018; Qian & Murphy, 2011; Swaminathan & Joachims, 2015; Zhao et al., 2012; Zhou et al., 2017). In particular, Athey & Wager (2017); Zhou et al. (2017), among others, highlight the importance of using *efficient* estimates of policy value as optimization objectives, *i.e.*, having minimal asymptotic mean-squared error (MSE). Examples of efficient estimators are direct modeling or IPW when outcome functions or propensities are sufficiently smooth (Hahn, 1998; Hirano et al., 2003), or DR leveraging cross-fitting (Chernozhukov et al., 2018) in more general non-parametric settings.

Regardless of which of these three estimates one uses, the resulting optimization problem amounts to a difficult binary optimization problem. Therefore many of the above leverage a reduction of this problem to weighted classification (for two actions; cost-sensitive classification more generally) and leverage tractable convex formulations that use classification surrogate loss functions for the zero-one loss, such as, for example, hinge loss (Zhao et al., 2012; Zhou et al., 2017, which yields a weighted SVM) and logistic loss (Jiang et al., 2019, which yields a weighted logistic regression). The recently proposed entropy learning approach of Jiang et al. (2019) is particularly appealing, since the logistic regression-based surrogate loss is smooth and therefore allows for statistical inference on the estimated optimal parameters. In general however, one may consider using any surrogate loss function that is *classification-calibrated* (Bartlett et al., 2006), meaning that any policy that minimizes the surrogate loss is optimal.

However, as we here emphasize, even if we use policy value estimates that are efficient, this *does not* imply that we obtain efficient estimation/learning of the optimal policy itself, *even* if the surrogate-loss model is well-specified. For example, in the case of logistic loss, we demonstrate that, although logistic regression is statistically efficient for actual

[1]Cornell University, and Cornell Tech, New York. Correspondence to: Andrew Bennett <awb222@cornell.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

binary classification when well-specified (as is well-known), in the case of policy learning via a weighted-classification reduction well-specification only implies a *semi*-parametric model, and therefore minimizing the empirical average of the surrogate loss is *not* efficient in this case.

On the other hand, the implications of correct specification can be summarized as a conditional moment problem. Such problems are amenable to efficient solution using approaches based on the generalized method of moments (GMM; Hansen, 1982). We demonstrate what an efficient such estimate would look like, in terms of the efficient instruments for our specific policy learning problem. We propose a particular implementation of solving our problem based on recent work on efficiently solving conditional moment problems using a reformulation of the efficient GMM solution as a smooth game optimization problem, which can be solved using adversarial training of neural networks (Bennett et al., 2019). In addition, we prove some results relating the efficiency of optimal policy estimation to the asymptotic regret of the surrogate loss, and also prove that under correct specification the regret of the surrogate loss upper bounds the true regret of policy learning.

We demonstrate empirically over a wide range of scenarios that our methodology indeed leads to greater efficiency, with lower MSE in estimating the optimal policy parameter estimates under correct specification. Furthermore, we demonstrate that in practice, both *with* and *without* correct specification, our methodology tends to learn policies with *lower regret*, particularly in the low-data regime.

1.1. Setting and Assumptions

Let X denote the context of an individual, $T \in \{-1, 1\}$ the treatment assigned to that individual, and Y the resultant outcome. In addition let $Y(t)$ denote the counterfactual outcome that would have been obtained for the corresponding individual if, possibly contrary to fact, treatment t had been assigned instead. We assume throughout that we have access to logged data consisting of n iid observations, $\mathcal{S}_n = \{(X_i, T_i, Y_i) : i \leq n\}$, of triplets (X, T, Y) generated by some behavior policy.

We make standard causal assumptions of consistency and non-interference, which can be summarized by assuming that $Y = Y(T)$. Furthermore, as is standard in the above policy learning literature, we assume that X encapsulates all possible confounders, that is, $Y(t) \perp T \mid X \; \forall t \in \{-1, 1\}$, as would for example be guaranteed if the logging policy is a function of the observed individual context.

A policy π denotes a mapping from individual context to treatment to be assigned. Concretely, given individual context x, let $\pi(x) \in \{-1, +1\}$ denote the treatment assigned by policy π (we may also consider stochastic policies but since optimal policies are deterministic we focus on these).

Let

$$\begin{aligned} J(\pi) &= \mathbb{E}[Y(\pi(X))] - \frac{1}{2}\mathbb{E}[Y(+1) + Y(-1)] \\ &= \mathbb{E}[\pi(X)(Y(+1) - Y(-1))] \end{aligned}$$

denote the expected value of following policy π, relative to complete randomization. Given the logged data and some policy class Π, our task is to learn an *optimal policy* from the class, defined by $\pi^* \in \arg\max_{\pi \in \Pi} J(\pi)$ (notice that offsetting by the complete randomization policy does not affect this optimization problem). In particular we consider policy classes where each policy π is indexed by some utility function g and is defined by $\pi(x) = \text{sign}(g(x))$, where in turn the utility functions are parametrized by $\theta \in \Theta \subseteq \mathbb{R}^d$ as $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$, so that

$$\Pi = \{\text{sign}(g_\theta(x)) : \theta \in \Theta\}.$$

Correspondingly, we define

$$J(\theta) = J(\text{sign}(g_\theta(\cdot))) = \mathbb{E}[\text{sign}(g_\theta(X))(Y(+1) - Y(-1))]$$

and $\theta^* \in \arg\max_{\theta \in \Theta} J(\theta)$. A prominent example is linear decision rules, where $g_\theta(x) = \theta^T x$. Other examples include decision trees of bounded depth and neural networks.

Unlike some past work that has considered non-parametric policy classes (*e.g.*Zhao et al. (2012)), which have advantages in terms of regret without having to rely on "correct specification" assumptions, we make the decision to focus on parametric policy classes only. This is because these classes are more amenable to efficiency analysis, and are very relevant in practice due to reasons such as interpretability and implementability.

1.2. Efficiency

We briefly review what it means to estimate the optimal policy parameters, θ^*, efficiently. For simplicity, suppose that θ^* is unique. A *model* $\mathcal{M}$ is some set of distributions for the data-generating process (DGP), *i.e.*, a set of probability distributions for the triplet (X, T, Y). A model is generally non-parametric in the sense that this set of distributions can be arbitrary, infinite, and infinite dimensional.

Consider any learned policy parameters $\hat{\theta}$, that is, a function of the data $\mathcal{S}_n$ with values in Θ. Roughly speaking, we say that $\hat{\theta}$ is *regular* if, whenever the data is generated from $(X_i, T_i, Y_i) \sim p \in \mathcal{M}$, we have that $\sqrt{n}(\hat{\theta} - \theta^*)$ converges in distribution to some limit as $n \to \infty$ and this limit holds in a particular locally uniform sense in $\mathcal{M}$ (see Van der Vaart, 2000, Chapter 25 for a precise definition). Semiparametric efficiency theory (see ibid.) then establishes that there exists a covariance matrix V such that for any cost function $c : \mathbb{R}^d \to \mathbb{R}$ for which the sublevel sets are

$\{v : c(v) \leq c_0\}$ are convex, symmetric about the origin, and closed, we have that

$$\liminf_{n\to\infty} \mathbb{E}[c(\sqrt{n}(\hat{\theta} - \theta^*))] \geq \mathbb{E}_{v\sim\mathcal{N}(0,V)}[c(v)] \quad (1)$$

for any estimator $\hat{\theta}$ that is regular in $\mathcal{M}$. An important example is MSE, given by $c(v) = \|v\|_2^2$.

Efficient estimators are those for which Eq. (1) holds with *equality* for all such functions c, which, by the portmanteau lemma, would be implied if the estimator has the limiting law $\sqrt{n}(\hat{\theta} - \theta^*) \Rightarrow \mathcal{N}(0, V)$. Regular estimators is a very general class of estimators so the bound in Eq. (1) is rather strong. So much so that, in fact, Eq. (1) holds in a local asymptotic minimax sense for *all* estimators (see ibid., Theorem 25.21).

Efficiency is important for practitioners because, in observational data, we only have the data that we have and cannot experiment or simulate to generate more so we should use the data optimally. Amongst other things, efficiency implies we can construct optimally-tight confidence intervals for the estimated optimal parameter values, and in Section 4.4 we show that efficiency implies asymptotically-optimal regret.

1.3. Related Work

There has been a variety of past work on the problem of policy learning from observational data. Much of this work considers formulating the objective of policy learning as a weighted classification problem (Beygelzimer & Langford, 2009; Dudík et al., 2011), and either minimizing the 0-1 loss directly using combinatorial optimization (Athey & Wager, 2017; Kitagawa & Tetenov, 2018; Zhou et al., 2018), using smooth stochastic policies to obtain a nonconvex but smooth loss surface (Swaminathan & Joachims, 2015), or replacing the 0-1 objective with a convex surrogate to be minimized instead (Beygelzimer & Langford, 2009; Dudík et al., 2011; Jiang et al., 2019; Zhao et al., 2012; Zhou et al., 2017). In addition there is work that extends some of the above approaches to the continuous action setting (Chernozhukov et al., 2019; Kallus & Zhou, 2018; Krishnamurthy et al., 2019); our focus will be solely on binary actions. Of these methods the convex-surrogate approach has the advantage of computational tractability and, when the convex surrogate is smooth (*e.g.* Jiang et al., 2019), the ability to perform statistical inference on the optimal parameters. Our paper extends this work by investigating how to solve the smooth surrogate problem efficiently. Although much of this past work has used objective functions for learning based on statistically efficient estimates of policy value (Athey & Wager, 2017; Chernozhukov et al., 2019; Dudík et al., 2011; Zhou et al., 2018), to the best of our knowledge our paper is novel in investigating the efficient estimation of the optimal policy parameters themselves.

In addition there has been a variety of past work on solving conditional moment problems (see Bennett et al. (2019); Khosravi et al. (2019) and citations therein). Our paper builds on this work as it reformulates the problem of policy learning as a conditional moment problem, which we propose to solve using optimally weighted GMM (Hansen, 1982) and DeepGMM (Bennett et al., 2019).

2. The Surrogate-Loss Reduction and Its Fisher Consistency

In this section, we present the surrogate-loss reduction of policy learning and the implications of correct specification.

Many policy learning methods start by recognizing that the policy value can be re-written as

$$J(\theta) = \mathbb{E}[\psi \operatorname{sign}(g_\theta(X))] \quad (2)$$

where ψ is any of the following score variables, which all depend on observables:

$$\begin{aligned} \psi_{\text{IPS}} &= \tfrac{TY}{e_T(X)}, \quad \psi_{\text{DM}} = \mu_1(X) - \mu_{-1}(X), \\ \psi_{\text{DR}} &= \psi_{\text{DM}} + \psi_{\text{IPS}} - \tfrac{T\mu_T(X)}{e_T(X)}, \end{aligned} \quad (3)$$

where $e_t(x) = P(T = t \mid X = x)$ and $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$. Equation (2) arises once we recognize that all of these satisfy $\mathbb{E}[\psi \mid X] = \mathbb{E}[Y(1) - Y(-1) \mid X]$.

Then we can approximate Eq. (2) using its empirical version:

$$J_n(\theta) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \psi_i \operatorname{sign}(g_\theta(X_i)). \quad (4)$$

In particular, Athey & Wager (2017); Kitagawa & Tetenov (2018); Zhou et al. (2018) prove bounds of the form $\sup_{\theta\in\Theta} |J_n(\theta) - J(\theta)| = O_p(1/\sqrt{n})$ given that the policy class has bounded complexity. This shows that optimizing $\hat{\theta} \in \arg\max_{\theta\in\Theta} J_n(\theta)$ provides near-optimal solutions in the original policy learning problem, since $J(\theta^*) - J(\hat{\theta}) \leq J(\theta^*) - J(\hat{\theta}) + J_n(\hat{\theta}) - J_n(\theta^*) \leq 2\sup_{\theta\in\Theta} |J_n(\theta) - J(\theta)|$. Given that in practice the nuisance functions e_t and μ_t are estimated from data, we denote the corresponding score variable when such estimates are plugged in as $\hat{\psi}$ to differentiate it from the variable ψ that uses the true nuisance functions. We correspondingly let $\hat{J}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \hat{\psi}_i \operatorname{sign}(g_\theta(X_i))$. When $\hat{J}_n(\theta)$ is efficient for $J(\theta)$ one can generally additionally prove that $\sup_{\theta\in\Theta} |\hat{J}_n(\theta) - J_n(\theta)| = o_p(1/\sqrt{n})$.

Given the non-convexity and non-smoothness of the empirical objective function Eq. (4), it is not necessarily clear how to actually optimize it. Many works (Beygelzimer & Langford, 2009; Jiang et al., 2019; Zhao et al., 2012) recognize that this optimization problem is actually equivalent to weighted binary classification (in our two-action case), since $\psi_i \operatorname{sign}(g_\theta(X_i)) = |\psi_i|(1 - 2\mathbb{I}_{\operatorname{sign}(g_\theta(X_i))\neq\psi_i})$, so any classification algorithm that accepts instance weights can

perhaps be used to address Eq. (4). Specifically, many classification algorithms take the form of minimizing a *convex surrogate loss*:

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} |\psi_i| l(g_\theta(X_i), \text{sign}(\psi_i)), \tag{5}$$

where $l(g, s)$ acts as a surrogate for the zero-one loss $\mathbb{I}_{\text{sign}(g_\theta(X_i))\neq\psi_i}$. For classification, Bartlett et al. (2006) studies which losses are appropriate surrogates, *i.e.*, are classification-calibrated. The population version of the surrogate loss, which $L_n(\theta)$ is approximating, is

$$L(\theta) = \mathbb{E}[|\psi| l(g_\theta(X), \text{sign}(\psi))]. \tag{6}$$

Following Jiang et al. (2019), we will focus on the logistic (or, logit-cross-entropy) loss function, which is defined according to

$$l(g, s) = 2\log(1 + \exp(g)) - (s + 1)g.$$

We note that although we focus on this loss in our discussions and experiments, all of our theoretical results, except for Lemma 1 which is specific to logistic regression loss, still hold with any other surrogate loss that satisfies the following assumptions.

Assumption 1. *The surrogate loss l is convex and twice differentiable in its first argument, and classification-calibrated.*

The *classification-calibrated* property is from Bartlett et al. (2006). We describe this property and justify that it is satisfied by logistic regression loss in the appendix. We also note that the other parts of Assumption 1 are clearly satisfied by logistic regression loss.

Furthermore, we note that some of our results can be extended to other convex, classification-calibrated losses that are not smooth, such as the hinge loss. We provide details of this in the appendix.

Given Assumption 1 and an additional regularity assumption, the following theorem follows immediately.

Assumption 2. $\mathbb{E}[\psi \mid X] = \mathbb{E}[Y(1) - Y(-1) \mid X]$, *and* $\mathbb{E}[|\psi|] < \infty$.

Theorem 1 (Fisher Consistency Under Correct Specification)**.** *Suppose the policy class Π is correctly specified for the surrogate loss in the sense that*

$$\mathcal{G} \cap \left(\underset{g \text{ unconstrained}}{\arg\min} \; \mathbb{E}[|\psi| l(g(X), \text{sign}(\psi))] \right) \neq \varnothing. \tag{7}$$

Then given Assumption 2, any minimizer of the surrogate-loss risk is an optimal policy:

$$J(\theta^*) = \max_{\pi \text{ unconstrained}} J(\pi)$$
$$\text{for all } \theta^* \in \arg\min_{\theta\in\Theta} L(\theta).$$

Theorem 1 establishes that, under correct specification, if we minimize the population surrogate loss, $L(\theta)$, then we obtain the optimal policy. Therefore, a natural strategy for policy learning would be to directly minimize the empirical loss $L_n(\theta)$, as was done by the above. Although the above arguments indicate that this approach would be computationally tractable, and also consistent under mild regularity conditions that ensure that optimizers of $L_n(\theta)$ would converge to optimizers of $L(\theta)$, it is not clear that it is statistically efficient, even if we use an efficient score variable for policy value estimation.

We provide a cautionary note that, as discussed *e.g.*in Qiu et al. (2019); Wager (2020), hoping for correct specification to hold for a simple low-dimensional policy class such as linear policies could be unreasonable. In such a case, it is possible that using this surrogate objective could systematically lead to incorrect policy decisions (Wager, 2020). On the other hand, if we use a flexible policy class such as neural networks of a given architecture, it is reasonable to assume that correct specification holds (at least approximately), so the surrogate objective may be better justified.

3. The Conditional-Moment Reformulation of the Surrogate-Loss Reduction

In this section we establish a new interpretation of the surrogate-loss reduction as a conditional moment problem and we discuss the implications of this in terms of the model implied by correct specification. This will enable us to conduct efficiency analysis and to design algorithms with improved efficiency in the next section.

3.1. The Conditional Moment Problem

First, we define l' as the derivative of l with respect to its first argument. In the case of logistic regression loss this is

$$l'(g, s) = 2\sigma(g) - (s + 1),$$

where $\sigma(g) = \exp(g)/(1 + \exp(g))$ is the logistic function.

Theorem 2 (Conditional Moment Problem Under Correct Specification)**.** *Suppose Assumption 2 holds and the policy class Π is correctly specified for the surrogate loss in the sense that Eq. (7) holds. Define*

$$m(X; \theta) = \mathbb{E}\left[|\psi| l'(g_\theta(x), \text{sign}(\psi)) \mid X\right].$$

Then we have that

$$\theta^* \in \arg\min_{\theta\in\Theta} L(\theta)$$
$$\iff m(X; \theta^*) = 0 \;\; \textit{almost surely}. \tag{8}$$

Theorem 2 arises straightforwardly from the observation that, under correct specification, $g_\theta(x)$ minimizes

$\mathbb{E}[|\psi_i| l(g_\theta(X), \text{sign}(\psi)) \mid X = x]$ for almost every x. Using smoothness and convexity, this latter observation is restated using first-order optimality conditions. Dominated convergence theorem allows us to exchange differentiation and expectation and we obtain the result. Theorem 2 provides an alternative characterization of θ^* as solving a *conditional moment problem*.

Notice that Eq. (8) is *equivalent* to the statement that, for any square integrable function f of X, we have the moment restriction

$$\mathbb{E}\left[m(X;\theta) f(X)\right] = 0\,.^{1} \tag{9}$$

This alternative characterization makes the problem amenable to efficiency analysis.

Notice that by first-order optimality, if $\theta^* \in \text{interior}(\Theta)$, optimizing $L(\theta)$ in Eq. (6) exactly corresponds to solving the set of d moment equations given by $\mathbb{E}[m(X;\theta)\nabla_\theta g_\theta(x)] = 0$. Similarly, optimizing the empirical loss $L_n(\theta)$ in Eq. (5) corresponds to solving these d equations with population averages ($\mathbb{E}$) replaced with empirical sample averages.

However, Eq. (9) gives a much broader set of equations. Leveraging this fact will be crucial to achieving efficiency. Indeed, it is well-known that even if a small number of moment equations are sufficient to identify a parameter (*e.g.*, in the above, the d equations identify θ^* via first-order optimality), taking into consideration additional moment equations that are known to hold can increase efficiency in semiparametric settings (Carrasco & Florens, 2014).

3.2. The Semiparametric Model Implied by Specification

In order to reason about efficiency, we need to reason about the model implied by Eq. (8). To do so, we first establish the following lemma.

Lemma 1. *Assume Assumption 2, and that we are using logistic regression loss. Then given a policy class* Π*, the model of DGPs (distributions on* (X,T,Y)*) where* Π *is correctly specified for the surrogate loss (in the sense of Eq.* (7)*) is given by all distributions on* (X,T,Y) *for which there exists* $\theta^* \in \Theta$ *satisfying*

$$\frac{\mathbb{E}[|\psi| \mathbb{1}\{\psi > 0\} \mid X = x]}{\mathbb{E}[|\psi| \mid X = x]} = \sigma(g_{\theta^*}(X)) \;\; \textit{almost surely}. \tag{10}$$

This model is generally a *semiparametric* model. That is, while Eq. (10) is a parametric restriction on the function $\mathbb{E}[|\psi|\mathbb{1}\{\psi > 0\} \mid X = x]/\mathbb{E}[|\psi| \mid X = x]$, the set of corresponding distributions on (X,T,Y) that satisfy this restriction is still infinite-dimensional and non-parametric.

[1] This is because in an inner product space $\mathcal{V}$, $v = 0$ if and only if $\langle v, v' \rangle = 0$ for every $v' \in \mathcal{V}$. Here $\mathcal{V}$ is L_2.

3.3. Comparison with Logistic Regression for Classification

One question the reader might have at this point is why an approach different than empirical loss minimization is necessary for efficiency, given that the surrogate loss formulation seems mathematically identical to binary classification using logistic regression, which is known to be efficient.[2] The difference between the problems is that for actual classification we have that ψ is a binary class label, *i.e.*, $\psi \in \{-1, 1\}$. If we assume the policy class is well-specified and $\psi \in \{-1, 1\}$, the characterization of our semiparametric model from Lemma 1 reduces to

$$P(\psi = 1 \mid X) = \sigma(g_{\theta^*}(X)),$$

which implies that our model is *parametric*, since the choice of θ^* now fully characterizes the distribution of the label ψ given X. *E.g.*, usually for logistic regression we let $g_\theta(x) = \theta^T x$ so that the above says that the logit of $P(\psi = 1 \mid X)$ is linear. Therefore, performing logistic regression corresponds to MLE for this parametric model, which is efficient.

However in our more general setting this is *not* the case and there is a non-trivial nuisance space, since there is a complex, infinite-dimensional space of conditional distributions for ψ given $X = x$ that could result in the *same* function $\mathbb{E}[|\psi|\mathbb{1}\{\psi > 0\} \mid X = x]/\mathbb{E}[|\psi| \mid X = x]$. This suggests that we may need to be more careful in order to obtain efficiency and that there may exist estimators that are more efficient than empirical loss minimization.

4. Efficient Policy Learning Reductions

In this section we propose some efficient methods for policy learning based on the above conditional-moment formulation. In addition, we provide some analysis of these methods in terms of efficiency and regret.

4.1. FiniteGMM Policy Learner

We begin by proposing an approach based on using multi-step GMM to solve the conditional moment problem, which we will call FiniteGMM. This approach works by optimally enforcing for the moment conditions given by Eq. (9) for a finite collection of critic functions $\mathcal{F} = \{f_1, \ldots, f_k\}$. Specifically, given some initial estimate $\tilde{\theta}_n$ of θ^*, define:

$$\begin{aligned} m(\theta)_j &= \tfrac{1}{n}\textstyle\sum_{i=1}^n |\hat{\psi}_i| l'(g_\theta(X_i), \text{sign}(\hat{\psi}_i)) f_j(X_i) \\ C(\tilde{\theta}_n)_{jk} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \hat{\psi}_i^2 l'(g_{\tilde{\theta}_n}(X_i), \text{sign}(\hat{\psi}_i))^2 f_j(X_i) f_k(X_i) \\ O(\theta; \tilde{\theta}_n) &= m(\theta)^T C(\tilde{\theta}_n) m(\theta). \end{aligned}$$

[2] This is because logistic regression performs maximum likelihood estimation (MLE), which is statistically efficient for well-specified parametric models.

We then estimate θ by $\hat{\theta}_n = \arg\min_\theta O(\theta; \tilde{\theta}_n)$. We can repeat this multiple times, plugging in $\hat{\theta}_n$ as $\tilde{\theta}_n$ and resolving.

An important issue with this estimator, however, is how to choose the critic functions. Standard GMM theory requires that the k moment conditions are sufficient to identify θ^*. And even then, the above is only the most efficient among estimators of the form $\arg\min_\theta \|(m_1(\theta), \dots, m_k(\theta)\|$ for any norm $\|\cdot\|$, but there may still be more efficient choices of critic functions.

4.2. The Efficient Instruments for Policy Learning

One nice result from the theory of conditional moment problems is the existence of a finite set of critic functions ensuring efficiency in the sense of Section 1.2. Define:

$$\begin{aligned}
\Omega(x) &= \mathbb{E}[\psi^2 l'(g_\theta(X), \text{sign}(\psi))^2 \mid X = x] \\
h_{\theta^*}(x) &= \nabla_\theta g_\theta(x) \mid_{\theta=\theta^*} \\
D(x) &= \mathbb{E}[\nabla_\theta(|\psi| l'(g_\theta(X), \text{sign}(\psi))) \mid_{\theta=\theta^*} \mid X = x] \\
&= \mathbb{E}[|\psi| l''(g_{\theta^*}(X), \text{sign}(\psi)) h_{\theta^*}(x) \mid X = x] \\
f_i^*(x) &= \frac{D(x)_i}{\Omega(x)}.
\end{aligned}$$

We call $\mathcal{F}^* = \{f_1^*, \dots, f_d^*\}$ the *efficient instruments*, and as long as the span of $\mathcal{F}$ contains these instruments then FINITEGMM is guaranteed efficiency (Newey, 1993). We can observe that these equations correspond to linearizing the moment equation at θ^*, and Ω is akin to the Fisher information matrix. We refer readers to Newey (1993) for more details on the derivation of these efficient instruments.

Given this, one approach would be to let $\mathcal{F}$ be flexible with the hope of approximately containing $\mathcal{F}^*$. Letting, for example, $\mathcal{F}$ be the first $k(n)$ functions in a basis for L_2 such as a polynomial basis and letting $k(n) \to \infty$ can be shown to be efficient under certain conditions (Newey, 1993). This, however, can perform very badly in practice, especially when the number of features is high, due to known curse of dimensionality issues with such classical nonparametric regression methods (Bauer et al., 2019; Geenens et al., 2011; Nagler & Czado, 2016).

4.3. ESPRM Policy Learner

Motivated by the above concerns, we now present our proposed approach: ESPRM (efficient surrogate policy risk minimization). This is based on the extension of Bennett et al. (2019) to our conditional moment problem. In the setting of instrumental variable regression, Bennett et al. (2019) proposes an adversarial reformulation of optimally-weighted GMM, which allows us to consider critic functions given by flexible classes such as neural networks. Then if this class provides a good approximation for the efficient instruments, this approach should be approximately efficient. Specifically, we define:

$$\begin{aligned}
u(X, \psi; \theta, f) &= |\psi| l'(g_\theta(X), \text{sign}(\psi)) f(X) \\
U(\theta, f; \tilde{\theta}) &= \tfrac{1}{n} \textstyle\sum_{i=1}^n u(X_i, \hat{\psi}_i; \theta, f)] \\
&\quad - \tfrac{1}{4n} \textstyle\sum_{i=1}^n u(X_i, \hat{\psi}_i; \tilde{\theta}_n, f)^2,
\end{aligned}$$

where as above $\tilde{\theta}_n$ is some initial consistent estimate of θ^*. Then following Bennett et al. (2019), the ESPRM estimator is defined as

$$\hat{\theta}^{\text{ESPRM}} = \arg\min_\theta \sup_{f \in \mathcal{F}} U(\theta, f; \tilde{\theta}),$$

where $\mathcal{F}$ is our flexible function class (henceforth assumed to be a class of neural networks). In brief, the motivation of this objective is as follows. First, if we let $\mathcal{F} = \text{span}(f_1, \dots, f_k\}$ in this objective, it follows from a generalization of Bennett et al. (2019, Lemma 1) that ESPRM is identical to FINITEGMM using critic functions $\{f_1, \dots, f_k\}$. It follows that ESPRM is equivalent to replacing the span of the critic functions in FINITEGMM with a generic function space. Therefore, instead of trying to approximate the efficient instrument using a growing basis for L_2 — an approach which is known to suffer from curse of dimensionality issues — we can instead use a flexible function space, such as a space of neural networks, that is designed to handle flexible function approximation without such issues (Bauer et al., 2019). We describe the theory behind this estimator in more detail in the appendix.

It remains to describe how this adversarial game is to be solved, and how to define $\tilde{\theta}_n$. As in Bennett et al. (2019) we optimize the objective by performing alternating first-order optimization steps using the OAdam algorithm (Daskalakis et al., 2017), which was designed for solving smooth game problems such as generative adversarial networks (GANs). In addition, we continuously update $\tilde{\theta}_n$ during optimization, where at each step of alternating first order optimization we set $\tilde{\theta}_n$ equal to the previous iterate of $\hat{\theta}_n$.

4.4. Efficient Learning implies Optimal Regret

Finally we prove that efficiency not only ensures minimal MSE in estimating θ^* but also implies regret bounds. Let

$$\begin{aligned}
\text{Regret}_J(\theta) &= \arg\max_{\pi \text{ unconstrained}} J(\pi) - J(\theta) \\
\text{Regret}_L(\theta) &= L(\theta) - \inf_{\theta \in \Theta} L(\theta).
\end{aligned}$$

Theorem 3 (Regret Upper Bound)**.** *Suppose Assumptions 1 and 2 hold and that the policy class Π is correctly specified for the surrogate loss in the sense that Eq. (7) holds. Then, for any $\theta \in \Theta$ we have*

$$\varphi(\textit{Regret}_J(\theta)) \leq \textit{Regret}_L(\theta)\,,$$

for some continuous φ that depends only on l, and satisfies $\varphi(0) = 0$ and $\varphi(\alpha) > 0$ for $\alpha > 0$. Furthermore, in the

case of logistic regression loss this bound reduces to

$$\textit{Regret}_J(\theta) \leq 2\textit{Regret}_L(\theta)\,.$$

This theorem implies that the regret of a policy is upper-bounded by the excess risk of the surrogate loss. Next, we make the following regularity assumption about the loss L:

Assumption 3 (Uniquely Minimized Loss)**.** *L has a unique minimizer θ^* in the interior of Θ.*

We note that this assumption is very strong, and may be unrealistic for classes such as neural networks where multiple different parameter values can correspond to the same function (for example by permuting the units in hidden layers). However, we argue that in practice this is not a major issue, and can be addressed for example by using symmetry-breaking constraints.

Given Assumptions 1 and 3, a Taylor's theorem expansion yields $\text{Regret}_L(\hat{\theta}_n) = (\hat{\theta}_n - \theta^*)^T H(\theta^*)(\hat{\theta}_n - \theta^*) + o(\|\hat{\theta}_n - \theta^*\|^2)$, where $H(\theta^*)$ is the Hessian of L at θ^*. For for any regular estimator $\hat{\theta}_n$, we can also define the *asymptotic regret* AR_L as the limiting distribution:

$$n\text{Regret}_L(\hat{\theta}_n) \to_d \text{AR}_L(\hat{\theta}_n),$$

which exists since regularity implies that $\sqrt{n}(\hat{\theta}_n - \theta^*)$ has a limiting distribution. Given this we can prove the following optimality result of our efficient estimators in terms of asymptotic regret:

Theorem 4 (Optimal Asymptotic Regret)**.** *Given Assumption 3 and any non-negative, non-decreasing ϕ, we define the risk $R_\phi(\hat{\theta}_n) = \mathbb{E}[\phi(AR_L(\hat{\theta}_n))]$. Given this, there exists a risk bound B_ϕ such that $R_\phi(\hat{\theta}_n) \geq B_\phi$ for every regular $\hat{\theta}_n$, with equality if $\hat{\theta}_n$ is semi-parametrically efficient.*

Together with Theorem 3, this means that both the actual regret (Regret_J) and the surrogate regret (Regret_L) of policies given by efficient estimators $\hat{\theta}$ are $O_p(1/n)$, and the surrogate regret has an optimal constant.

Note that this is not the first regret result for policy learning based on convex surrogate losses. Of particular interest, Zhao et al. (2012) previously provided optimal regret results for nonparametric policy learning using hinge loss. However, unlike us they showed that their regret obtains an optimal rate but *not* optimal constants, and their result depends on various additional technical assumptions.

5. Experiments

5.1. Synthetic Scenarios

First we investigate the performance of our algorithms on a variety of synthetic scenarios, using logistic regression loss for l. In all these scenarios X is 2-dimensional, and X and $Y(t) - \mu_t(X)$ are standard Gaussian distributed for each t; the scenarios only differ in the functions μ_t and e_t. In none of the scenarios is our policy class *actually* well-specified in the sense of Eq. (7).

We consider the following kinds of synthetic scenarios:

- LINEAR: $\mu_t(x) = a_t^T x + a_{t0}$ and $e_1(x) = \text{sigmoid}(b^T x + b_0)$ for some vectors a_{-1}, a_1, b.
- QUADRATIC: $\mu_t(x) = x^T A_t x + a_t^T x + a_{t0}$ and $e_1(x) = \text{sigmoid}(x^T B x + b^T x + b_0)$ for some symmetric matrices A and B, and vectors a_{-1}, a_1, b.

In addition we experiment with the following policy classes: a *linear* policy class, where $g_\theta(x) = \theta^T x + \theta_0$, and a *flexible* policy class where $g_\theta(x)$ is given by a fully-connected neural network with a single hidden layer of size 50, and leaky ReLU activations.

In all cases we use the surrogate loss method of Jiang et al. (2019) described in Section 2 as a benchmark, which we henceforth refer to as ERM. We note that although in the prior work they used $\hat{\psi}_{\text{IPS}}$, we instead use $\hat{\psi}_{\text{DR}}$, both because it is theoretically better grounded given its double robustness property (Athey & Wager, 2017; Zhou et al., 2017) and we found that it gives stronger results for all methods. For our ESPRM method we let $\mathcal{F}$ be the same neural network function class as for flexible policies, and perform alternating first-order optimization as described in Section 4.3 for a fixed number of epochs. For FINITEGMM we experimented with function sets $\mathcal{F}$ based on various polynomial basis expansions, and also various finite-dimensional approximations of Gaussian kernel basis expansions using the method of Random Kitchen Sinks (Rahimi & Recht, 2009). We provide details of these function sets in the appendix.

For all methods, except where otherwise specified, we use the $\hat{\psi}_{DR}$ weights described in Eq. (3), with nuisance functions fit using correctly specified linear regression or logistic regression algorithms on a separately sampled tuning dataset of the same size as the training dataset.[3] We provide some additional results in the appendix where nuisances were instead fit via flexible neural networks, which show that this has little effect on our results. In all cases except for ESPRM we perform optimization using LBFGS. Additional optimization details are given in the appendix.

For all configurations of scenario kind and policy we ran our experiments by sampling random scenarios of the respective kind, by setting all scenario parameters to be independent standard Gaussian variables. Specifically, for each

[3] By correctly specified we mean that for LINEAR we fit using linear/logistic regression on X, whereas for QUADRATIC we fit on a quadratic feature expansion of X.

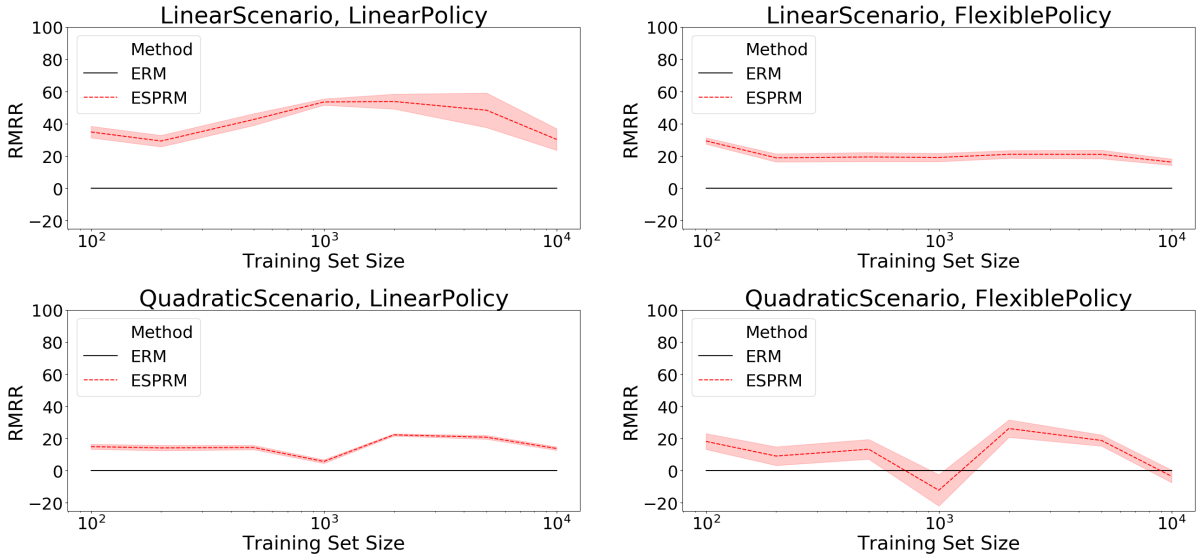

Figure 1. Difference in performance between ESPRM and ERM. We plot RMRR against training set size for each combination of policy class and scenario kind. Shaded regions are 95% confidence intervals, constructed from bootstrapping using the 64 replications.

$n \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$ we sample 64 random scenarios of the respective kind, and for each random scenario we sample n training data points and run all methods on this data. Results for FINITEGMM, which generally did badly as predicted for all the basis function sets described above, are given in the appendix.

Define Relative Mean Regret Reduction (RMRR), given by:

$$RMRR(\hat{\theta}_n) = \left(1 - \frac{\mathbb{E}[\text{Regret}_J(\hat{\theta}_n)]}{\mathbb{E}[\text{Regret}_J(\hat{\theta}_n^{\text{ERM}})]}\right) \times 100\%,$$

where each expectation in the fraction is taken over the joint distribution of randomly sampled scenarios, and the corresponding random estimates $\hat{\theta}$. Then for each scenario kind and policy class, we plot predicted RMRR against number of training data based on our ESPRM estimates in Fig. 1. In addition, we provide plots for this experiment on a raw utility scale in the appendix. Note that the confidence intervals in all these plots for each data point are with respect to the joint distribution over randomly generated scenarios and corresponding random estimate $\hat{\theta}$.

We see that ESPRM consistently obtains policies on average that are lower regret or on-par than those obtained by ERM (typically with around 10% to 20% RMRR), with the 95% confidence intervals indicating clearly better performance in almost every case, except for in the case of training flexible policies in the quadratic scenario where there are a couple of outlier points, although even there the two methods seem at worst roughly on-par. We can also observe that the most significant regret benefits tend to occur with smaller training set sizes (since the same RMRR implies a larger absolute decrease in regret), indicating that the statistical efficiency of our method is leading to improved finite sample behavior.

In Fig. 2 we plot the convergence in terms of the MSE of the estimated parameter from ESPRM and ERM, for the LINEAR setting and linear policy class (where parameters are low-dimensional and correctly specified). We plot both the MSE convergence, and the average difference in the squared error between the estimates, across the random scenarios.[4] It is clear from these results that ESPRM consistently estimates optimal policy parameters with lower squared error on average compared to ERM across these random simulated scenarios. This provides strong evidence that the methodology indeed provides an improvement in statistical efficiency for solving the smooth surrogate loss problem.

5.2. Jobs Case Study

We next consider an application to a dataset derived from a large scale experiment comparing different programs offered to unemployed individuals in France (Behaghel et al., 2014). We focus our attention to two arms from the experiment: a treatment arm where individuals receive an intensive counseling program run by a public agency and a treatment arm with a similar program run by a private agency. The hypothetical application is learning a personalized allocation to counseling program, with the aim of maximizing the number of individuals who reenter employment within six months, minus costs. (The original study's focus was not

[4] All parameter vectors are normalized first given that the policy function is scale-invariant.

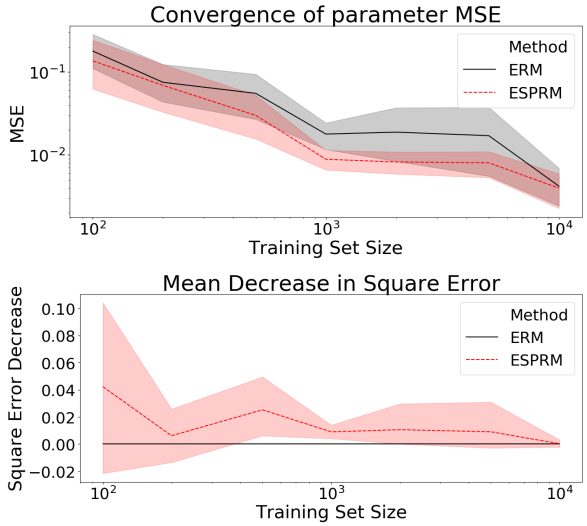

Figure 2. Above we plot the convergence in MSE of the predicted $\hat{\theta}_n$ for each method with a linear policy class, over the random scenarios of the LINEAR class. Below we plot the average difference in the squared error of ESPRM and ERM (positive numbers indicate improvement over ERM). All shaded regions are 95% confidence intervals constructed from bootstrapping.

personalization.) Our intervention is simply the offer of the counseling program; we therefore ignore the fact that some individuals offered one of the programs did not attend.

In order to make our policies focus on heterogeneous effects, so that a constant-treatment policy wouldn't be optimal and policy learning would be non-trivial, we set the costs of each arm to be equal to their within-arm average outcome in the original data. That is, the outcome we consider is equal whether one reentered employment within 6 months, minus the average number of individuals who entered employment within 6 months in that arm, so therefore each arm has a mean outcome of zero. The covariates we consider personalizing on are: statistical risk of long-term unemployment, whether individual is seeking full-time employment, whether individual lives in sensitive suburban area, whether individual has a college education, the number of years of experience in the desired job, and the nature of the desired job (*e.g.*, technician, skilled clerical worker, etc.).

We then consider 64 replications of the following procedure. Each time, we randomly split the data 40%/60% into train/test. We then introduce some confounding into the training dataset. We consider the following three binary variables: whether individual has 1–5 years experience in the desired job, whether they seek a skilled blue collar job, and whether their statistical risk of long-term unemployment is medium. After studentizing each variable, we segment the data by the tertiles of their sum. In the first tertile, we drop each unit with probability $7/8$. In the second tertile, we drop private-program units with probability $1/4$ and public-program units with probability $7/8$. In the third tertile, we drop public-program units with probability $1/4$ and private-program units with probability $7/8$. Given a policy learned on this training data, we evaluate it on the held-out test set using a Horvitz-Thompson estimator.

Policy	ERM	ESPRM	Difference
Linear	-0.96 ± 4.32	4.42 ± 3.78	5.38 ± 5.06
Flexible	-1.75 ± 4.64	7.68 ± 3.16	9.42 ± 5.17

Table 1. Average predicted policy value (multiplied by 1000) for the Jobs case study for ERM versus ESPRM over 64 repetitions. The $\pm$ interval provides the 95% confidence intervals.

Of the training data, 20% was set aside for training nuisances, and an additional 20% as validation data for early stopping. We then trained both linear and flexible policies using ERM and ESPRM as in our simulation studies, with the exception that nuisances were fitted using neural networks (of the same architecture as the flexible policy class).

We summarize the mean estimated outcome for the policies from each method in Table 1. We note from these values that on average ESPRM seems to be learning higher value job-assignment policies than ERM. In addition, we conducted paired two-sided t-tests to test the hypothesis that the two algorithms lead to different mean policy values on this data, under the randomness in our data splitting and confounding procedures as well as the estimation algorithms. We obtained p-values of .0429 for the linear policy class and .0007 for the flexible policy class, clearly highlighting the benefit of our ESPRM method.

6. Conclusion

We considered a common reduction of learning individualized treatment rules from observational data to weighted surrogate risk minimization. We showed that, quite differently from actual classification problems, assuming correct specification in the policy learning case actually suggests more efficient solutions to this reduction. In particular, even if we use efficient policy evaluation, this may not necessarily lead to efficient policy learning. Specifically, under correct specification, the problem becomes a conditional moment problem in a semiparametric model and efficiency here translates to both better MSE in estimating optimal policy parameters and improved regret bounds.

Based on this observation, we proposed an algorithm, ESPRM, for efficiently solving the surrogate loss problem. We showed that our method consistently outperformed the standard method of empirical risk minimization on the surrogate loss, both over a wide variety of synthetic scenarios and in a case study based on a real job training experiment.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1846210.

References

Athey, S. and Wager, S. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Bauer, B., Kohler, M., et al. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics*, 47(4):2261–2285, 2019.

Behaghel, L., Crépon, B., and Gurgand, M. Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *American economic journal: applied economics*, 6(4):142–74, 2014.

Bennett, A., Kallus, N., and Schnabel, T. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, pp. 3559–3569, 2019.

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

Carrasco, M. and Florens, J.-P. On the asymptotic efficiency of gmm. *Econometric Theory*, 30(2):372–406, 2014.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Chernozhukov, V., Demirer, M., Lewis, G., and Syrgkanis, V. Semi-parametric efficient policy learning with continuous actions. In *Advances in Neural Information Processing Systems*, pp. 15039–15049, 2019.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Geenens, G. et al. Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys*, 5:30–43, 2011.

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

Hansen, L. P. Large sample properties of generalized method of moments estimators. *Econometrica*, pp. 1029–1054, 1982.

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

Jiang, B., Song, R., Li, J., and Zeng, D. Entropy learning for dynamic treatment regimes. *Statistica Sinica*, 2019.

Kallus, N. Recursive partitioning for personalization using observational data. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1789–1798, 2017.

Kallus, N. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pp. 8895–8906, 2018.

Kallus, N. and Zhou, A. Policy evaluation and optimization with continuous treatments. *arXiv preprint arXiv:1802.06037*, 2018.

Khosravi, K., Lewis, G., and Syrgkanis, V. Non-parametric inference adaptive to intrinsic dimension. *arXiv preprint arXiv:1901.03719*, 2019.

Kitagawa, T. and Tetenov, A. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Krishnamurthy, A., Langford, J., Slivkins, A., and Zhang, C. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *arXiv preprint arXiv:1902.01520*, 2019.

Nagler, T. and Czado, C. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

Newey, W. K. Efficient estimation of models with conditional moment restrictions. *Handbook of Statistics*, 11: 419–454, 1993.

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

Qiu, H., Luedtke, A., and Laan, M. Comment on 'entropy learning for dynamic treatment regimes' by binyan jiang, rui song, et al. *Statistica Sinica*, (29), 2019. doi: 10.5705/ss.202019.0062.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pp. 1313–1320, 2009.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pp. 814–823, 2015.

Van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, 2000.

Wager, S. On regression tables for policy learning: comment on a paper by jiang, song, li and zeng. *Statistica Sinica*, (29), 2020. doi: 10.5705/ss.202019.0071.

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

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

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.