Policy Evaluation and Optimization with Continuous Treatments

Nathan Kallus
Cornell University and Cornell Tech

Angela Zhou
Cornell University

Abstract

We study the problem of policy evaluation and learning from batched contextual bandit data when treatments are continuous, going beyond previous work on discrete treatments. Previous work for discrete treatment/action spaces focuses on inverse probability weighting (IPW) and doubly robust (DR) methods that use a rejection sampling approach for evaluation and the equivalent weighted classification problem for learning. In the continuous setting, this reduction fails as we would almost surely reject all observations. To tackle the case of continuous treatments, we extend the IPW and DR approaches to the continuous setting using a kernel function that leverages treatment proximity to attenuate discrete rejection. Our policy estimator is consistent and we characterize the optimal bandwidth. The resulting *continuous policy optimizer* (CPO) approach using our estimator achieves convergent regret and approaches the best-in-class policy for learnable policy classes. We demonstrate that the estimator performs well and, in particular, outperforms a discretization-based benchmark. We further study the performance of our policy optimizer in a case study on personalized dosing based on a dataset of Warfarin patients, their covariates, and final therapeutic doses. Our learned policy outperforms benchmarks and nears the oracle-best linear policy.

1 Introduction

Personalization is a key feature of modern decision making in a variety of contexts and learning how to personalize is a central problem in machine learning. In targeted advertising, learning systems observe data about incoming users such as their market segment, decide which ad to display to the individual user, and observe whether or not the user clicks on the ad. In personalized medicine, medical history, demographics, and genetics of patients may be leveraged to administer individually tailored treatments. Contextual-bandit algorithms, where the learning system repeatedly takes an action for a context (feature vector) and observes the corresponding outcome, and other randomized experiments such as A/B tests, are the gold standards for comparing the efficacy of different policies and learning the best one. However, in practice, it can be prohibitively costly, risky, or impossible to collect new data through experimentation. Off-policy evaluation and optimization is the problem of assessing and optimizing new personalized decision policies based on observational data collected under other historical policies.

The existing literature on policy evaluation has primarily considered only discrete action spaces, where the learning system chooses one of K treatments for each unit [3, 4, 16], except for model-based approaches for policy evaluation [12]. In many important applications, however, the treatment is a continuous variable. For example, the dosage of a medical drug is continuous and, by using personalized dosing policies, doctors may adjust dosages to account for individual factors such as genes. As another example, in dynamic pricing, different values of customer rebates (e.g., from 10% to 40%) can be viewed as continuous treatments offered to the customer. The duration of or intensity of exposure to an intervention, such as a job training program, can be considered as a continuous treatment as well. Treating such variables as discrete, e.g., by discretizing the data, must rely on ad-hoc modeling and may impede the fidelity of evaluation and the performance of optimized policies. Interpreting treatments as continuous is helpful even if not all continuum values are observed since such an interpretation allows off-policy evaluation to learn across treatments that are different but close.

We present a framework for policy evaluation and op-

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

timization with continuous treatments. Our proposed estimator, introduced in Sec. 2, effectively uses outcome data from data points where the treatment was close to the target policy. In Sec. 3, we analyze the bias, variance, and the mean-squared error of our policy evaluation. For the corresponding estimated optimal policy, we analyze the consistency of policy optimization in Sec. 4. Finally, in Sec. 5, we conduct experiments showing that our approach performs well. In particular, we consider a case study with clinical data from a dataset of warfarin patients.

2 Policy Evaluation Methodology

2.1 Problem Setup and Notation

We study an observational dataset collected from interactions of the decision-making system assigning treatments to units x_i. For each interaction, the system observes an input (context or feature) $x_i \in \mathcal{X}$. Following the logging policy, the system assigns some treatment $t_i \in \mathcal{T}$ with probability Q_i. Outcome data y_i is observed which is generated from a joint distribution on the covariates, treatments, and induced outcomes, $(X, T, \{Y(t)\}_{t \in \mathcal{T}}) \sim \mathcal{D}$, unknown to us. The observational dataset comprises of n i.i.d. observations of data $(x_i, t_i, y_i(t_i))) \sim \mathcal{D}$.

The term *treatment* corresponds to *arms* or *actions* discussed in other works on off-policy evaluation. The generalized propensity score (GPS) is defined as $Q_i = f_{T|X}(T = t_i, X = x_i) = f_{T|X}(t_i \mid x_i)$ and generalizes the discrete propensity score $p_a = \mathbb{P}[t_i = a]$ for the continuous setting; we assume it exists [7]. We assume the logging policy is known, which is reasonable when we have control over the system. Otherwise, the GPS can be imputed with standard approaches for predicting conditional densities such as regression under parametric noise models or kernel density estimation.

Off-policy evaluation estimates the policy value

$$V_\tau = \mathbb{E}[Y(\tau(X))]$$

the value of the expected outcomes induced by the policy $\tau(X)$, corresponding to potential outcomes under the Neyman-Rubin causal framework. $Y(t)$ denotes the potential outcome of a unit had it received treatment t [19]. We require standard assumptions in causal inference of *unconfoundedness* (also known as ignorability) and *common support*. Unconfoundedness asserts that $Y_i(t) \perp T_i \mid X_i$ for all t: treatment is exogenous and its assignment depends only on x_i. The data generation process described above is consistent with unconfoundedness. The *common support* condition requires that $f_{T|X}(\tau(x_i) \mid x_i) > 0$: otherwise, if possible treatments may never be observed in the dataset, there is no chance for accurate estimation of their respective outcomes.

The task of policy optimization is to determine the optimal policy within a restricted function class of policies $\mathscr{T}$. Since the optimal policy is deterministic (for each user, assign the optimal treatment), we focus on evaluating deterministic policies. The empirically optimal policy $\hat{\tau} \in \operatorname{argmin}_{\tau \in \mathscr{T}} \hat{v}^n_\tau$ is the policy minimizing the estimated value from our proposed estimator. The best-in-class policy $\tau^* \in \operatorname{argmin}_{\tau \in \mathscr{T}} V_\tau$ minimizes the unknown expected policy value.

2.2 Related Work

Previous approaches for off-policy evaluation broadly include the direct method (DM), inverse propensity weighted estimators (IPW), or estimators which combine them. The direct method estimates the relationship between outcomes and the union of covariates x and treatment t, $r(X, T) = \mathbb{E}[Y \mid X, T]$, and generates a plug-in estimator. By unconfoundedness, regression-based estimation of the conditional mean function $\mathbb{E}[\mathbb{E}[Y|X, T = \tau(X)]]$ corresponds to estimation of the potential outcome [19]. However, this approach is subject to issues with model misspecification and without addressing dataset imbalance from a logging policy, may over or under-estimate the relevance of outcomes under a different policy [3]. In [12], the authors have investigated the evaluation of continuous treatment effects with the Super-Learner, an ensemble model which incorporates multiple models of the entire dose-response surface.

IPW-based estimation normalizes the observed outcomes by the inverse of the *propensity weights* of the logging policy [8]. IPW estimation corrects distribution mismatch by averaging outcomes over a new dataset created out of reweighted instances where the logging and target policies assign the same treatment [16]. IPW is unbiased, with a slower rate of convergence dependent on the number of treatments [3], and it is optimal in the sense of minimax efficiency when no additional information about the reward structure is available [24]. However, dividing by the propensity score can inflate the variance of IPW estimators.

The doubly-robust (DR) estimator combines DM and IPW estimators. When the direct estimate of the reward estimator is biased, such as when using non-parametric or high-dimensional regression of $\mathbb{E}[Y \mid X, T]$, the doubly robust estimator weights the model residuals by inverse propensity weights in order to remove the bias. DR achieves a multiplicative bias when propensities are estimated and its convergence requires only that one of the estimators are consistent [23, 4]. Recent work in [23, 24] switches between using an IPS and reward estimator, using the reward estimator when the propensity is smaller than some threshold which optimizes the MSE bias-variance trade-off. In [21, 22], the authors propose counterfactual risk minimization for policy optimization by minimizing an upper bound on the MSE of an IPW-based estimator.

The continuous setting for policy evaluation and optimization presents new challenges. We note that the generalized propensity score as introduced in [7] is analyzed in the context of treatment effect evaluation, and is used in practice to motivate appropriate discretizations of a continuous treatment variable for assessing balance. Policy optimization in discrete action spaces is generally reduced to a weighted multi-class classification problem, where the classes are treatments and are weighted by their off-policy evaluation. For each context, the policy determines the action which provides highest rewards as its classification label [4]. However, policy optimization in the continuous setting will be fundamentally different since the problem does not decompose into discrete classes of outcomes.

2.3 Off-Policy Continuous Estimator

Previous IPW approaches for off-policy evaluation in discrete action spaces propose the following estimator, which filters the observational dataset by rejection and importance sampling:

$$\mathbb{E}\left[\tfrac{1}{n}\textstyle\sum_i \tfrac{y_i}{Q_i}\mathbb{I}\{\tau_i(x_i)=t_i\}\right]$$

In the continuous setting, we will not be able to employ rejection sampling since $\mathbb{P}[\tau(x_i)=t_i]=0$ for any continuous probability density. The rejection sampling term $\mathbb{I}\{\tau(x_i)=t_i\}$ can be viewed as a Dirac delta function, $\delta_{\tau(x_i)}(t_i)$, and in the discrete case, it enforces that the only outcome data used for estimation are the observations where the same treatment was observed under the logging policy and is assigned by the target policy. For continuous treatments, our proposed estimator re-weights the dataset to consider outcomes where the observed treatment and off-policy treatment are close.

We propose the continuous-treatment off-policy evaluator, denoted as $\hat{v}_\tau$, which smoothly relaxes the unit mass of the Dirac delta function using a kernel function $K(u)$:

$$\hat{v}_\tau = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tau(x_i)-t_i}{h}\right)\frac{y_i}{Q_i}$$

Properties of the kernel function $K(u)$ include symmetry about the origin ($\int uK(u)du=0$) and that it integrates to 1 ($\int K(u)du=1$). Kernel density estimates, also known as Parzen-window estimation, can be viewed as smooth nonparametric generalizations of computing histogram 'buckets'. Instead of assuming a specific parametric statistical model, kernel density estimation assumes smoothness of the underlying joint density [6]. We state our results for univariate kernels where $\mathcal{T}\subseteq\mathbb{R}$ and note that analogous results hold if we use multidimensional kernel functions. When $\mathcal{T}\subseteq\mathbb{R}^d$, the estimator takes the form $\hat{v}_\tau=\frac{1}{n}\sum_{i=1}^n |H|^{-\frac{1}{2}}K(|H|^{-\frac{1}{2}}(\tau(x_i)-t_i))\frac{y_i}{Q_i}$, where H denotes a bandwidth matrix. Examples of kernels include Gaussian kernels, where $K(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$ or the Epanechnikov kernel, $K(u)=\frac{3}{4}(1-u^2)\mathbb{I}\{|u|\leq 1\}$.

This approach extends the IPW and rejection sampling approach taken in discrete treatment spaces to continuous treatment spaces. The extent of the kernel smoothing, parametrized by the bandwidth h, can be chosen to minimize the mean-squared-error (MSE). In particular, our estimator differs from using the direct method with kernel-based regression of the conditional density $f_{Y|T,X}$ and evaluation of the estimate at the treatment policy. Thus, we avoid the curse of dimensionality: kernel regression performs dramatically worse as the number of covariate dimensions increases, whereas the convergence rates of our method rely only on the treatment dimension [17].

In the special case that the logging policy is unknown such that propensities Q_i must be imputed, and if a dose-response estimator $\hat{r}(t,x)$ is available, our estimator can be extended to a doubly robust one that has bias in excess of that in Thm. 1 that is multiplicative in the estimation biases of propensities and dose response:

$$v_\tau^{\mathrm{DR}} = \\ \tfrac{1}{n}\textstyle\sum_{i=1}^n \left[\hat{r}(\tau(x_i),x_i)+\tfrac{1}{hQ_i}K\left(\tfrac{\tau(x_i)-t_i}{h}\right)(y_i-\hat{r}(t_i,x_i))\right]$$

2.3.1 Self-Normalized Propensity Weight estimator

As discussed in [22], IPW methods can suffer from variance in estimates due to the propensity weights. Normalizing the IPW estimator by $\sum {}^1/_{Q_i}K(\frac{\tau(x_i)-t_i}{h})$ maintains consistency but can reduce variance by adjusting estimates of the treatment space that would have been be sampled with greater or lower probability.

$$\hat{v}_\tau^{\mathrm{norm}} = \frac{\sum_{i=1}^n \frac{y_i}{Q_i}K\left(\frac{\tau(x_i)-t_i}{h}\right)}{\sum \frac{1}{Q_i}K\left(\frac{\tau(x_i)-t_i}{h}\right)}$$

2.3.2 Practical Concerns

When implementing our off-policy evaluation estimator in practice, some adjustments need to be made for empirical performance.

Bandwidth selection: Selecting a good bandwidth is key to good evaluation and optimization. We compute the asymptotically optimal bandwidth in Thm. 1 below, but beyond the order in n, the expression includes constants that are generally unknown a priori. In the case of kernel density estimation, the literature focuses on methods for bandwidth selection which do not incorporate loss scalings and would perform poorly in our setting [18]. Instead, we propose to select the optimal bandwidth via a plug-in estimator, estimating the quantities in the expression for optimal bandwidth (eq. 3.2): we estimate the conditional density via kernel density estimation and subsequently estimate the second derivative and the conditional expectation via numerical integration.

Boundary bias: If the treatment space is bounded, the kernel may extend past the boundaries where necessarily no data point exists, biasing boundary estimates downwards. This can be addressed by truncating and normalizing the kernel: if $\mathcal{T} = [T_{\text{lo}}, T_{\text{hi}}]$ then we simply divide each term in our estimator by $\int_{T_{\text{lo}}}^{T_{\text{hi}}} K((\tau(x_i) - t)/h)dt$.

Clipping propensity weights: When using IPW-based estimators, in practice if the propensity score is very small, it is clipped with some threshold θ, e.g., 0.1. This introduces additional bias but may significantly reduce the variance, yielding smaller total error.

3 Off-Policy Evaluation Analysis

3.1 Bias and Variance of Kernelized Policy Evaluation

We compute the bias and variance of the estimator $\hat{v}_\tau$ and prove consistency. Some technical assumptions are required for the analysis:

Assumption 1. *The conditional outcome and treatment densities, $f_{Y|T,X}$ and $f_{T|X}$, exist.*

Assumption 2. *The conditional outcome density $f_{Y|T,X}$ is twice differentiable and the conditional treatment density $f_{T|X}$ is differentiable.*

Assumption 3. *Outcomes y_i are bounded with finite second moments.*

Assumption 4. *Common support between the treatment propensities observed in the data and the treatment policy $\tau(x_i)$: $f_{T|X}(\tau(x_i), x) \geq a > 0$, for almost everywhere x and some fixed a.*

Assumption 5. *(Unconfoundedness) $Y_i(t) \perp T_i \mid X_i$ for all t: treatment is exogenous and its assignment depends only on x_i.*

Theorem 1. *Under Assumptions 1-5, the bias and variance of $\hat{v}_\tau$ are:*

$$\text{Bias}(\tau) := \mathbb{E}\left[\hat{v}_\tau - V_\tau\right] =$$

$$\kappa_2(K)\mathbb{E}\left[\int \frac{y_i}{2}\frac{\partial^2}{\partial T^2} f_{Y|T,x}(y_i, \tau(x_i))dy\right] h^2 + o(h^2)$$

$$\text{Var}(\tau) := \mathbb{E}[(\hat{v}_\tau - \mathbb{E}[\hat{v}_\tau])^2] =$$

$$R(K)\mathbb{E}\left[\frac{\mathbb{E}[Y^2 \mid \tau(X), X]}{f_{T|X}(\tau(X), X)}\right]\frac{1}{nh} + O(\frac{h^4}{n}) + o\left(\frac{1}{nh}\right)$$

where $\kappa_2(K) = \int u^2 K(u)du$ and $R(K) = \int K(u)^2 du$ are the second moment and roughness of the kernel, respectively.

Proof outline. The theorem follows by applying Bayes' rule with the GPS and Taylor expansion of $f_{Y|T,X}$ around $\tau(x_i)$. Details in Appendix 7.1. □

The bias introduced by kernel density estimation is $O(h^2)$ and depends on the curvature of the unknown density $f_{Y|T,X}$ evaluated at the policy $\tau(X_i)$: if the outcome distribution changes rapidly with small changes in treatment value, our approach for leveraging local information will incur more bias. The variance depends inversely on the generalized propensity score. As expected, the estimator may have high variance in regions where we are unlikely to observe treatment $\tau(x_i)$.

3.2 Mean Squared Error and Consistency

We analyze mean squared error derived from bias and variance and characterize the optimal bandwidth. Intuitively, the bandwidth controls the scale of proximity we require on treatments: a bandwidth too large introduces high bias because we simply average over the entire dataset, while small bandwidths increase variance.

Theorem 2. *Under Assumptions 1-5, the MSE of $\hat{v}_\tau$ is*

$$\mathbb{E}\left[(\hat{v}_\tau - V_\tau)^2\right] =$$

$$\frac{R(K)}{nh}\mathbb{E}\left[\frac{\mathbb{E}[Y^2|\tau(X),X]}{f_{T|X}(\tau(X)|X)}\right] + O(h^4) + O\left(\frac{h^4}{n}\right) + o\left(\frac{1}{nh}\right)$$

and the optimal bandwidth h^* is $\Theta(n^{-\frac{1}{5}})$.

$$h^* = \left(\frac{R(K)\mathbb{E}\left[\mathbb{E}[Y^2|\tau(X),X]/f_{T|X}(\tau(X),X)\right]}{4(\mathbb{E}\left[\int \frac{y_i}{2}\frac{\partial}{\partial T^2} f_{Y|T,x}(y_i,\tau(x_i))\kappa_2(K)dy\right])^2 n}\right)^{\frac{1}{5}}$$

Proof outline. The theorem follows from the bias-variance decomposition of MSE and using Theorem 1, then optimizing over h. (Appendix 7.2) □

Theorem 3. *Under Assumptions 1-5, if $1/(nh) \to 0$ then $\hat{v}_\tau$ is consistent for V_τ:*

$$\frac{1}{nh}\sum_{i=1}^n K\left(\frac{\tau(x_i)-t_i}{h}\right)\frac{y_i}{Q_i} \to_p V_\tau$$

Proof Outline. Follows from convergent MSE and Markov's inequality. Full proof in Appendix 7.3 □

Corollary 4. *Under Assumptions 1-5, if $1/(nh) \to 0$ then the self-normalized off-policy evaluation estimator is consistent for V_τ.*

$$\hat{v}_\tau^{\text{norm}} = \frac{\sum_{i=1}^n \frac{y_i}{Q_i} K\left(\frac{\tau(x_i)-t_i}{h}\right)}{\sum \frac{1}{Q_i} K\left(\frac{\tau(x_i)-t_i}{h}\right)} \to_p V_\tau$$

Proof. The result follows from Slutsky's theorem since $\frac{1}{nh}\sum_{i=1}^n \frac{1}{Q_i} K\left(\frac{\tau(x_i)-t_i}{h}\right) \to_p 1$. □

4 Continuous Policy Optimization

Accurate off-policy evaluation is a necessary prerequisite for policy optimization, the task of estimating which treatment policy minimizes expected desired outcomes. We analyze how the empirically optimal policy, the policy minimizing the off-policy evaluations, performs out-of-sample.

For a constrained policy class, such as a space of linear policies ($\mathscr{T} = \{\tau(x) = \beta^\intercal x : \|\beta\|_2 \leq W_2\}$), the policy optimization problem can be interpreted as a weighted empirical risk minimization problem over a constrained policy space where we find $\hat{\tau} \in \underset{\tau\in\mathscr{T}}{\operatorname{argmin}} \frac{1}{nh}\sum_{i=1}^n K\left(\frac{\tau(x_i)-t_i}{h}\right)\frac{y_i}{Q_i}$. Gradients can be computed easily with respect to the kernel function, applying the chain rule to $\tau(x)$, and we provide additional examples in the Appendix. Equivalently we can optimize the other estimators $\hat{v}_\tau^{\mathrm{norm}}, \hat{v}_\tau^{\mathrm{DR}}$ and incorporate variance regularization. The nonconvex optimization can be solved by a numerical optimizer such as L-BFGS or gradient descent, but generally with no guarantees for global convergence. In practice we take the best solution from random restarts.

4.1 Consistency of Policy Optimization

Our analysis bounds the generalization error for the empirical risk-minimizing policy, the error incurred by minimizing the empirical risk instead of the unknown expected risk. Generalization bounds for this problem follow from [1, Thm. 8], and depend on the Rademacher complexity of the loss function class. The empirical Rademacher and Rademacher complexity of a function class $\mathscr{T}$ are, respectively, defined as:

$$\hat{\mathcal{R}}_n(\mathscr{T}) = \mathbb{E}\left[\sup_{f\in\mathscr{T}}\left|\tfrac{2}{n}\sum_{i=1}^n \sigma_i f(x_i)\right| \mid x_1, \ldots x_n\right]$$
$$\mathcal{R}_n(\mathscr{T}) = \mathbb{E}[\hat{\mathcal{R}}_n(\mathscr{T})]$$

where σ_i are iid Rademacher random variables, symmetrically 1 or -1 with probability $\frac{1}{2}$. Restricting the function class provides better generalization error by reducing the Rademacher complexity: a function class which is less able to fit arbitrary data sequences is less vulnerable to overfitting.

Assumption 6. *Outcome values y_i are bounded on the interval $[-\bar{M}_y, \bar{M}_y]$. The inverse propensity weight, $\frac{1}{Q_i}$, is bounded on $[1, \bar{M}_Q]$.*

Assumption 7. *The kernel function $K(u)$ is bounded by $\bar{M}_K$ and has Lipschitz constant L_K.*

Theorem 5. *Under Assumptions 1-7, for any integer n number of samples and any $0 < \delta < 1$, with probability at least $1-\delta$, every $\tau \in \mathscr{T}$ satisfies:*

$$|V_\tau - \hat{v}_\tau| \leq \frac{L_K \bar{M}_Q \bar{M}_y}{h^2}\mathcal{R}_n(\mathscr{T}) + \frac{\bar{M}_y \bar{M}_Q \bar{M}_K}{h}\sqrt{\frac{2\log(2/\delta)}{n}} + |\mathrm{Bias}(\tau)|$$

Proof Outline. The result follows from the Rademacher generalization bound [1, Thm. 8] concentrating $\mathbb{E}\hat{v}_\tau$ near $\mathbb{E}\hat{v}_\tau$, Thm. 1 relating V_τ to $\mathbb{E}\hat{v}_\tau$, and the Rademacher comparison lemma [14, Thm. 4.12]. The full proof is provided in Appendix 7.5. □

Corollary 6. *Let $\hat{\tau} \in \operatorname{argmin}_{\tau\in\mathscr{T}} \hat{v}_\tau$ and $\tau^* \in \operatorname{argmin}_{\tau\in\mathscr{T}} V_\tau$. Then, under Assumptions 1-7, with high probability the regret of $\hat{\tau}$ satisfies*

$$V_{\hat{\tau}} - V_{\tau^*} \leq O_p(\tfrac{1}{h^2}\mathcal{R}_n(\mathscr{T}) + \tfrac{1}{h\sqrt{n}} + h^2)$$

The corollary shows that the regret of our policy optimizer converges to zero, i.e., achieves best-in-class performance, as long as $h = o(1)$, $h = \omega(\sqrt{\mathcal{R}_n(\mathscr{T})})$, and $h = \omega(1/\sqrt{n})$. As an example, consider a function class of linear decision rules with bounded norm: $\mathscr{T} = \{\tau(x) = \beta^\intercal x : \|\beta\|_2 \leq W_2\}$. Assuming that $\|X_i\|_2 \leq X_2$, [10] shows that the Rademacher complexity of this class is bounded as $\hat{\mathcal{R}}_n(\mathscr{T}) \leq \frac{W_2 X_2}{\sqrt{n}}$. Therefore, the optimal bandwidth $h = \Theta(n^{-\frac{1}{5}})$ ensures consistent policy learning of the best linear policy. Similar results hold for policies in a bounded ball of a reproducing kernel Hilbert space.

Variance regularization: If the optimization space includes a policy which assigns treatments arbitrarily far from the observed treatments, such a policy trivially minimizes the loss by forcing $K(\frac{\tau(x_i)-T_i}{h}) \to 0$. Regularizing the objective by the estimated sample standard deviation $\frac{1}{n}\sqrt{\sum_{i=1}^n (K(\frac{\tau(x_i)-t_i}{h})\frac{y_i}{Q_i} - \hat{v}_\tau)^2}$ of the policy evaluation should mitigate this effect from overly expressive policy classes.

5 Experiments

5.1 Validation on Synthetic Data

We first consider a controlled setting with synthetic data to illustrate our method. We consider $y = 2|x-t|^{1.5} + 0.2\epsilon$, where $\epsilon \sim N(0,1)$ and $x \sim \mathrm{Unif}[0,1]$. We consider treatment assignment that is either completely randomized uniformly on the interval $[-0.5, 1.3]$ without regard to x or treatment assignment that is confounded by x and is normally distributed as $T \mid X \sim N(x+0.1, 0.5)$. The optimal policy is linear and sets $\tau(x_i) = \beta x_i$ where $\beta = 1$.

We consider how the performance of off-policy evaluation changes with n by evaluating the optimal policy with n observational data points generated either using completely randomized treatments or treatments confounded by x, clipping generalized propensities below 0.1. We use the Epanechnikov kernel with the self-normalized estimator and estimate the bandwidth by using kernel density estimation of $f_{Y,T,X}$ and $f_{T,X}$ for the conditional density $f_{Y|T,X} = \frac{f_{Y,T,X}}{f_{T,X}}$. From this estimate we obtain an estimate for the second derivative $\frac{\partial^2}{\partial T^2} f_{Y|T,X}$ by numerical differentiation and compute an approximate conditional expectation of $\mathbb{E}[Y_i^2 \mid \tau(x_i), X]$ via numerical integration. Since computing the bandwidth is numerically intensive, we compute it for one value of n_0 and adjust it for different n_i by multiplying by $(\frac{n_0}{n_i})^{1/5}$. We compare to standard clipped-IPW discrete-treatment policy evaluation by discretizing the treatments into 10 evenly sized bins from the minimal to maximal observed treatment, computing the discrete propensity score by integrating the GPS over the bin ("discretized OPE"). We also compare against the direct method, using either a trained random forest regressor ("DM RF") or polynomial regression of order 3 ("DM Poly").

For each n between 10 and 300, we simulate 50 replications of the process. The results in Fig. 1 include 95% confidence intervals around the mean over replications. In both settings our policy evaluation indeed converges to the truth and while the discretization approach performs reasonably well, it is systematically biased, inconsistent, and has larger variance. The discretization is sensitive to the distribution of the data and the variation in the unknown true relationship between covariates, treatment, and outcome. In practical settings with real data, it is unclear what the best discretization would be.

To consider off-policy optimization in this simple setting, we fix $n = 300$ and evaluate linear policies with β ranging over $[0, 1.3]$. Again we consider 50 replications. Figs. 2 and 3 show 95% confidence intervals around the mean over replications. Our policy evaluations are tight near the optimum and subsequent optimization over the range of β values is consistent with the true optimal β. In Fig. 4 we evaluate the consistency of policy optimization by analyzing the out-of-sample error of the empirically optimal policy computed on datasets of n varying from 10 to 300. We compute 20 replications and display 95% confidence intervals around the mean, observing that following the theory, the out-of-sample error from off-policy evaluation converges to zero as n increases.

5.2 Policy Optimization Simulation

We consider a similar controlled setting in higher dimension with richer dose-response structure, still with synthetic data, and illustrate the resulting outcome distribution under various learned policies. We randomly generate independent ten-dimensional covariates ($d = 10$), normally distributed with zero mean and randomly generated covariances (following a normal distribution which is offset for positivity). The true outcome model is quadratic in T: we set the noiseless outcome as $y_i = \beta_T^\intercal T \beta_x^\intercal x_i + \beta_{x,T}^\intercal x_i T_i + (T_i - \beta_{x,T^2}^\intercal x)^2$ where $\beta_x \sim N(0, I_d)$, $\beta_{x,T} \sim N(0, 1.5 I_d)$, and $\beta_{x,T^2} \sim N(0, I_d)$. We induce sparsity on the coefficients by independently and randomly sampling 3 covariates to remain positive on each coefficient vector β_x, $\beta_{x,T}$, and β_{x,T^2}. We include a constant treatment effect interaction term of $\beta_T = -5$.

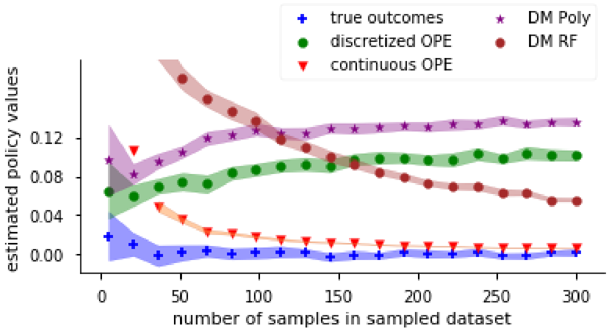

Figure 1: Off-policy evaluations and 95% confidence intervals as n increases, evaluating a linear treatment policy where $\beta = 1$ with normal sampling (model in Sec. 5.1).

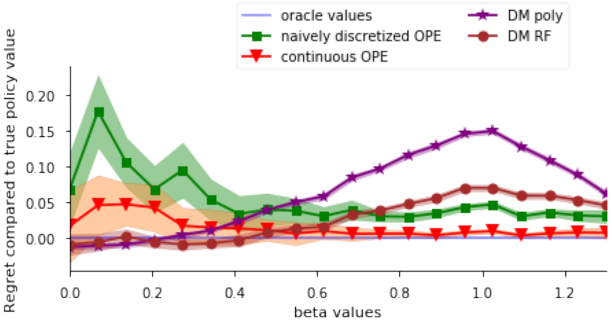

Figure 2: Regret of evaluations of a linear treatment policy compared to the true value, n=300, over different values of β with completely randomized sampling.

We sample treatments as normally distributed conditional on covariates, $T \sim N(\theta^T x, 4) + 2x_1 + 4x_2 - 3x_3$ and generate a training dataset of 400 instances (x_i, y_i, T_i) and evaluate on a test set of 1000 instances. Outcomes are generated from the model y_i with additional i.i.d. mean-zero Gaussian noise with variance 5. For policy learning, we consider the case that the propensity model is well-specified but unknown and impute the generalized propensity score from the training data via linear regression.

In Fig. 5 we compare the outcome distributions under learned policies using a box plot, displaying the means on the right. For reference we compute the best treatment assignment for each x_i given the full counterfactual model (best out-of-class (o.o.c)). We evaluate continuous off-policy evaluation over linear policies with a bandwidth of 2.6. When optimizing the off-policy

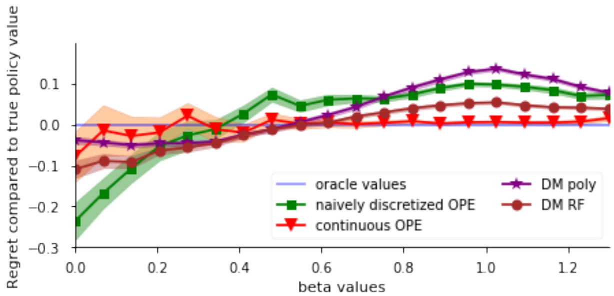

Figure 3: Regret of evaluations of a linear treatment policy compared to the true value, n=300, over different values of β with confounded sampling.

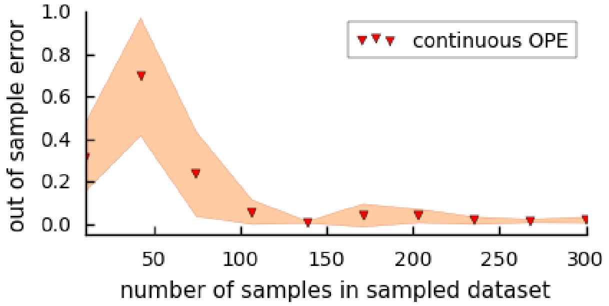

Figure 4: Out-of-sample error of empirically optimal policy from off-policy evaluation as n increases.

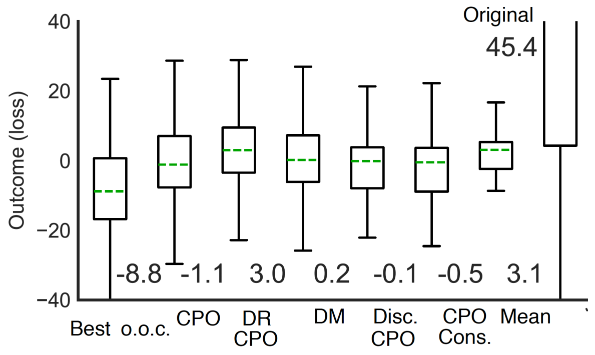

Figure 5: Boxplot of the distributions of outcomes for policy learning on 10-dimensional covariates with a quadratic outcome model (Sec. 5.2), displaying mean loss values.

estimator for the linear policy, we use the L-BFGS algorithm with multiple random restarts since the objective function is non-convex and L-BFGS performs well even in the non-convex setting [15]. We evaluate the direct method (DM), using a random forest regressor with a linear policy space, which we optimize by using the numerical differentiation available with BFGS. We use the same random forest regressor for doubly-robust continuous policy optimization (DR CPO). We also consider a discretization approach which optimizes over a continuous treatment policy (linear) but evaluates performance by discretizing treatment into 10 uniformly sized bins, running standard self-normalized CPE. Discretizing the resulting linear policy yields a constant policy in this setting: we compare to the best constant policy found using continuous policy evaluation (CPE, cons.). Finally, the baseline is a constant policy which assigns the mean dose. Comparing the results, we see that off policy evaluation is able to improve upon the mean risk of other methods and nears the performance of the best treatment assignment with full information (which is out of the linear policy class). While the best constant policy found using OPE has good performance in the sense of mean risk, the linear policy is better able to personalize treatment based on covariates.

5.3 Warfarin case study

Unlike for discrete off-policy evaluation, no evaluation datasets are available for continuous treatments with full counterfactuals. We evaluate our estimator in an experimental setting by developing a case study from a PharmGKB [9] dataset on warfarin dosing which includes information on patient covariates, final therapeutic dosages, and patient outcomes (INR, International Normalized Ratio). Warfarin is a blood thinner whose therapeutic dosage varies widely across patients and whose administration must be closely monitored to prevent adverse side effects. Previous work on predicting dosage policies [2, 11] has evaluated accuracy based on prediction of the correct category of dosage, "low" ($<$21 mg/wk), "medium" ($>$ 21 mg/wk,$<$ 49 mg/wk) or "high" ($>$ 49 mg/wk). However, clinical guidelines suggest fine adjustments to dosage (15-20%) during monitoring, and recommend splitting tablets to deliver precise treatment [9]. Therefore, treating warfarin dosage as a continuous variable better captures the inherently continuous nature of dosage amounts.

We develop a semi-simulated study by simulating a dosage process in a way that allows us to simulate counterfactual outcomes. Following the procedure set out in [9], we consider correct prediction as being within 10% of the therapeutic dose T_i^*, since measurements of patient INR are inherently noisy and dose is adjusted until the patient INR presents within a target range. Since the clinical risk of incorrect dosage increases with absolute distance from the target range [5], we use a semi-simulated loss function of absolute distance from $[.9T_i^*, 1.1T_i^*]$, instead of simulating unavailable INR outcomes:

$$y(\tau(x_i)) = \max(|\tau(x_i) - T_i^*| - 0.1T_i^*, 0)$$

We sample our dosage data T_i as a mixture of a patient's BMI z-score $Z_{\text{BMI}} = \frac{x_{\text{BMI}} - \mu_{\text{BMI}}}{\sigma_{\text{BMI}}}$ and i.i.d standard normal noise ϵ_i, scaled to preserve the moments of the therapeutic dose distribution, μ_T^*, σ_T^*, such that

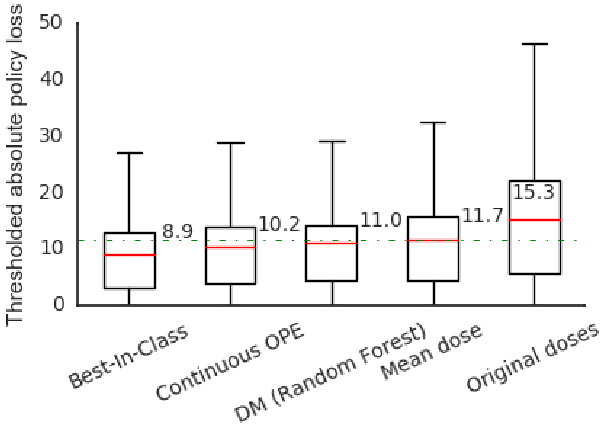

Figure 6: Boxplot of the distributions of thresholded absolute distances between policy and therapeutic doses, including mean loss values, for Warfarin study in Sec. 5.3.

Policy	Mean L1	Std. dev. L1	Mean L2
Best	8.93	8.64	154.37
Cont. OPE	10.19	10.19	207.78
DM	11.02	10.96	241.68
Mean dose	11.67	10.52	246.80
Original	15.27	13.08	404.28

Figure 7: Table of summary statistics of differences between estimated policy and true therapeutic doses

with $\theta = 0.5$: $T_i = \mu_T^* + \sigma_T^* \sqrt{\theta} Z_{\text{BMI}} + \sigma_T^* \sqrt{(1-\theta)}\epsilon$. It follows that the propensity score is $f_{T|X}(T_i' = t \mid x_{\text{BMI}}) = f_Z\left(\frac{t - \mu_T^* - \sigma_T^* \sqrt{\theta} Z_{\text{BMI}}}{\sqrt{1-\theta}}\right)$ where $\epsilon \sim N(0,1)$ and f_Z is the continuous density of a standard normal random variable.

We impose bounds on the coefficient, $\beta_d \in [-\frac{T_{\max}}{.25 d \mu_{X_d}}, \frac{T_{\max}}{.25 d \mu_{X_d}}]$, to prevent evaluating a policy with no overlap with the observed dataset, where $T_{\max}$ is the maximal treatment, μ_{X_d} is the mean of the dth covariate and d denotes dimension. We run a priori feature selection on the full dataset before evaluating policy optimization, using the importance weights from a random forest regressor on the therapeutic dose to select the 81 most important features.

We conduct policy optimization on these simulated outcomes and evaluate how the empirically optimal policy $\hat{\tau}_n$ performs on the thresholded loss function with absolute and squared penalties. The best-in-class linear treatment policy from median regression, $\tau^* \in \operatorname{argmin}_\tau \mathbb{E}[|\tau(x_i) - T_i^*| \mid X, T]$, has access to information about the true therapeutic dose ("Best-in-class" on the figure). We also evaluate the best linear model from a random forest regressor (DM estimator) for $\hat{r}(t_i, x_i)$. We compare against the linear policy found using our CPO method ("Continuous OPE") which achieves a mean loss of 10.2. The baseline is a constant policy corresponding to the mean dose and for reference we plot the distribution of outcomes according to the original initial treatment assignment observed in the dataset, which doctors adjusted until a therapeutic dose was reached when patient INR was within the target range. We tested discrete off-policy optimization (POEM and NORM-POEM) with various uniformly-spaced discretizations or quantile-based discretizations of dosages, but the propensity scores are mostly zero or one and hamper the resulting optimization, illustrating the difficulty of finding appropriate discretizations for real datasets [21, 22].

Comparing the results in Fig. 6, we see that our method is competitive with the best-in-class linear policy and improves upon the direct method, further reducing the median (Mean L1) and mean of the difference between the policy dose and therapeutic dose from the naive benchmark policy giving the mean dose. In Table 7 we also report the squared distance from $[0.9T_i^*, 1.1T_i^*]$ (mean L2): we see that performance of the DM policy shows less improvement from giving the mean dose when we weight outliers more heavily, and results in larger variance in the distribution of absolute losses (std. dev L1). Our approach for policy optimization based on continuous off-policy evaluation works well in simulated and semi-experimental settings.

6 Conclusion

We developed an inverse-propensity-weighted estimator for off-policy evaluation and learning with continuous treatments, extending previous methods which have only considered discrete actions. The estimator replaces the rejection sampling used in IPW-based estimators with a kernel function to incorporate local information about similar treatments. Our generalization bound for policy optimization shows that the empirically optimal treatment policy computed by minimizing the off-policy evaluation also converges to the policy minimizing the expected loss. We demonstrate the efficacy of our approach for estimation and evaluation on simulated data, as well as on a real-world dataset of Warfarin dosages for patients.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1656996. Angela Zhou is supported through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

[1] Peter Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

[2] Hamsa Bastani and Mohsen Bayati. Online decision-making with high-dimensional covariates. *Management Science*, 2015.

[3] Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[4] Miroslav Dudik, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014.

[5] Valentin Fuster, Lars E. Ryden, Davis S. Cannom, Harry J. Crijns, Anne B. Curtis, Kenneth A. Ellenbogen, Jonathan L. Halperin, G. Neal Kay, Jean-Yves Le Huezey, James E. Lowe, S. Bertil Olsson, Eric N. Prystowsky, Juan Luis Tamargo, and L. Samuel Wann. 2011 accf/aha/hrs focused updates incorporated into the acc/aha/esc 2006 guidelines for the management of patients with atrial fibrillation. *Circulation*, 2011.

[6] Bruce Hansen. Lecture notes on nonparametrics. Technical report, University of Wisconsin, 2009.

[7] Keisuke Hirano and Guido Imbens. *The Propensity Score with Continuous Treatments, in Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, chapter 7. John Wiley & Sons, Ltd, 2004.

[8] Daniel Horvitz and Donovan Thompson. A generalization of sampling without replacement froma finite universe. *Journal of the American Statistical Association*, 1952.

[9] T E International Warfarin Pharmacogenetics Consortium, Klein, R B Altman, N Eriksson, B F Gage, S E Kimmel, M-T M Lee, N A Limdi, D Page, D M Roden, M J Wagner, M D Caldwell, and Johnson J A. Estimation of the warfarin dose with clnical and pharmacogenetic data. *The New England Journal of Medicine*, 2009.

[10] Sham Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear predicttion: Risk bounds, margin bounds, and regularization. *Advances in Neural Information Processing Systems*, 2009.

[11] Nathan Kallus. Recursive partitioning for personalization using observation data. *Proceedings of the Thirty-fourth International Conference on Machine Learning*, 2017.

[12] Noemi Kreif, Richard Grieve, Ivan Dia, and David Harrison. Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury. *Health Economics*, 2015.

[13] Gert R. Lanckriet and Bharath K. Sriperumbudur. On the convergence of the concave-convex procedure. *Advances in Neural Information Processing Systems 22*, 2009.

[14] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

[15] Dong-Hui Li and Masao Fukushima. On the global convergence of bfgs method for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 2000.

[16] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. *Proceedings of the fourth ACM international conference on web search and data mining*, 2011.

[17] Adrian Pagan and Aman Ullah. *Nonparametric Econometrics*. Cambridge University Press, 1999.

[18] Byeong Park and J.S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 2009.

[19] Donald Rubin. Estimating causal eeffect of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974.

[20] Bernard Silverman. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 1978.

[21] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization. *Journal of Machine Learning Research*, 2015.

[22] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *Proceedings of NIPS*, 2015.

[23] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. *Journal of Machine Learning Research*, 2016.

[24] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. *Proceedings of Neural Information Processing Systems 2017*, 2017.