Identification of Subgroups With Similar Benefits in Off-Policy Policy Evaluation

Ramtin Keramati KERAMATI@CS.STANFORD.EDU
Stanford University, U.S.A

Omer Gottesman OGOTTESM@CS.BROWN.EDU
Brown University, U.S.A

Leo Anthony Celi LCELI@MIT.EDU
Massachusetts Institute of Technology, U.S.A

Finale Doshi-Velez FINALE@SEAS.HARVARD.EDU
Harvard University, U.S.A

Emma Brunskill EBRUN@CS.STANFORD.EDU
Stanford University, U.S.A

Abstract

Off-policy policy evaluation methods for sequential decision making can be used to help identify if a proposed decision policy is better than a current baseline policy. However, a new decision policy may be better than a baseline policy for some individuals but not others. This has motivated a push towards personalization and accurate per-state estimates of heterogeneous treatment effects (HTEs). Given the limited data present in many important applications such as health care, individual predictions can come at a cost to accuracy and confidence in such predictions. We develop a method to balance the need for personalization with confident predictions by identifying subgroups where it is possible to confidently estimate the expected difference in a new decision policy relative to a baseline. We propose a novel loss function that accounts for the uncertainty during the subgroup partitioning phase. In experiments, we show that our method can be used to form accurate predictions of HTEs where other methods struggle.

Data and Code Availability This research paper uses the MIMIC-III dataset (Johnson et al., 2016). This dataset is available on the PhysioNet repository (Goldberger et al., 2000). Code to generate results in section 6 (experiments) are available in the supplementary materials.

1. Introduction

Recent advances in technology and regulations around them have enabled the collection of an unprecedented amount of data of past decisions and outcomes in different domains such as health care, recommendation systems, and education. This offers a unique opportunity to learn better decision-making policies using observational data. Off-policy policy evaluation (OPE) is concerned with estimating the value of a proposed policy (*evaluation policy*) using the data collected under a different policy (*behaviour policy*).Estimating the value of an evaluation policy before deployment is essential, especially when interacting with the environment is expensive, risky, or unethical, such as in health care (Gottesman et al., 2019). Fortunately, the reinforcement learning (RL) community has developed different methods and theories focused on OPE, e.g. Jiang and Li (2016); Thomas and Brunskill (2016); Kallus and Uehara (2020).

OPE has been used extensively in the literature to demonstrate the superiority of a proposed evaluation policy relative to the baseline (behaviour) policy e.g. Komorowski et al. (2018); however, the evaluation policy may be better than the behaviour policy for some individuals but not others. Hence, only looking at the estimated value of the evaluation policy before deployment, may be misleading. In the non-sequential setting, a growing literature has focused on personalization and estimation of heterogeneous treatment effect (HTE), the individual-level

© 2022 R. Keramati, O. Gottesman, L.A. Celi, F. Doshi-Velez & E. Brunskill.

differences in potential outcomes under the proposed evaluation policy versus the behaviour policy (Athey et al., 2019; Nie and Wager, 2017).

In this paper, we aim to provide actionable information to domain experts. Specifically, we ask "*What subgroups of individuals can we confidently predict that will be significantly benefited or harmed by adopting the evaluation policy?*". Asking this question instead of "*What is the treatment effect for each individual?*" allows us to group individuals that have similar treatment effects together. When data is limited and the horizon of decision-making is long, estimates of individual treatment effects can have high variance resulting in high uncertainty. This means that it can be hard to assess whether a new policy will be more effective for a particular individual, rendering the goal of providing effective personalization for individuals unrealistic. Instead, here we provide an adaptive method to pool the data, to provide predictions that are both more accurate and confident, and can be used by a domain expert (human-in-the-loop) to make informed decisions before deploying the RL system or be used for the interpretability of the OPE. For example, a clinician can take a look at the groups and decide if the predicted benefit or harm is in accordance with their clinical intuition.

Prior work (Athey and Imbens, 2016) previously proposed a loss function that could be used to recursively partition the covariate space into groups in a way to balance variance and personalization for a single decision (non-sequential settings). We propose a similar approach for the sequential setting, we find that this prior loss function can be too noisy and often results in over-splitting, yielding too many subgroups and inaccurate or uncertain predictions. We mitigate the issue of noisy estimation by proposing a novel upper bound to the loss function that is stable and can be efficiently calculated. Our approach is relevant to conditions including intensive care treatment, cancer treatment, IVF, and physical therapy, which involve several stages and are often recorded in electronic medical records.

Additionally, by taking into account what clinicians believe is relevant for decision making, we incorporate a regularization term that prioritizes partitioning the space of individuals to create subgroups with treatment effects that exceed a specific (positive or negative) sufficient threshold. This threshold can be specified by a clinician and allows for better incorporation of medical experts into the evaluation process. For example, a clinician may consider an increase of 10% in survival rate relevant, so only subgroups with a confident prediction of 10% decrease or increases in survival rate will provide actionable information. Combining these two additions, we propose a new loss function.

On a simulated example of sepsis management (Oberst and Sontag, 2019), we show how our proposed method can be used to find subgroups with significant treatment effects, providing more accurate and confident predictions than related work that was developed for single-decision settings Athey and Imbens (2016). Additionally, we apply our method to the sepsis cohort of the MIMIC III ICU dataset (Johnson et al., 2016), and illustrate how it can be used to identify subgroups in which a new decision policy may be beneficial or harmful relative to the standard approach. We also investigate the interpretability of our findings through a discussion with an ICU intensivist.

2. Related Work

The need to estimate the value of a new decision policy is present in many different applications, such as personalized medicine (Obermeyer and Emanuel, 2016), bandits (Dudík et al., 2011) and sequential decision makings (Thomas and Brunskill, 2016). The RL community has developed different methods and theories for off-policy policy evaluation (OPE) in sequential settings. These methods mostly fall into different categories: importance sampling (Precup, 2000), model-based and doubly robust methods (Dudík et al., 2011; Thomas and Brunskill, 2016; Jiang and Li, 2016). All these methods can be used along with our algorithm to estimate group treatment effect for a particular group; however, these methods do not readily provide a way to perform partitioning into groups with distinct effects.

In non-sequential settings, a growing body of literature seeks to estimate heterogeneous treatment effect (HTE) using different approaches. For example, Imai and Ratkovic (2014) uses the LASSO to estimates the effect of treatments, Shalit et al. (2017) uses neural networks and offers a generalization bound for the individual treatment effect (ITE). Lee et al. (2020) relies partly on good confidence intervals to propose a robust partitioning method for non-sequential settings; whereas we perform upper bound variance estimation to avoid issues with very small effective sample sizes and noisy confidence intervals which plague sequential settings.

Our work draws a close parallel to methods using recursive partitioning to estimate HTEs (Athey and Imbens, 2016; Athey et al., 2019), but those works suffer from over-splitting the feature space in sequential settings due to noisy estimation of the loss function. We propose a different loss function that can be better estimated when data is limited, and that also incorporates domain expert knowledge regarding treatment effects that are of sufficient magnitude to be of interest.

3. Setting and Background

We consider an episodic stochastic decision processes with a finite action space $\mathcal{A}$, continuous state space $\mathcal{X} \in \mathbb{R}^M$, reward function $R : \mathcal{X} \times \mathcal{A} \to [0, R_{max}]$ and the discount factor $\gamma \in [0, 1]$. A policy π maps the state space to a probability distribution over the action space, and we assume each episode lasts at most H steps. A set of trajectories $\mathcal{T} = \{\tau_1, \ldots, \tau_N\}$ is provided. Each trajectory τ_i consists of a state x_t, action a_t and the observed reward r_t at step t, $\tau_i = \{x_0^i, a_0^i, r_0^i, \ldots, x_H^i\}$. Actions are generated by following a known behaviour policy π_b. We denote the evaluation policy by π_e.

4. Framework for Subgroup Identification

Our main goal is to robustly quantify the expected benefit or cost of switching from a behavior policy to a proposed evaluation policy for subsets of the population. To do so it is helpful to extend the standard notion of the treatment effect to the (sequential decision) policy treatment effect. We define the individual treatment effect $t(x; \pi_e, \pi_b)$ for a possible initial state x as

$$t(x; \pi_e, \pi_b) = \mathbb{E}_{\pi_e}\left[\sum_{t=0}^{H} \gamma^t r_t | x_0 = x\right] - \mathbb{E}_{\pi_b}\left[\sum_{t=0}^{H} \gamma^t r_t | x_0 = x\right]. \tag{1}$$

Before we introduce our definition of group treatment effects, we first define a partitioning over the state space by $L = \{l_1, \ldots, l_M\} \in \Pi$, such that $\bigcup_{i=1}^{M} l_i = \mathcal{X}$ and $\forall i, j : \quad l_i \bigcap l_j = \emptyset$. Define the partition function $l(x; L) = l_i$ such that $x \in l_i$. Given a partitioning L, partition-value function for a policy π can be defined as:

$$v(x; L, \pi) = \mathop{\mathbb{E}}_{\substack{x' \sim \mathcal{X} \\ a \sim \pi(.|x')}} \left[\sum_{t=0}^{H} \gamma^t r_t | x_0 = x', x' \in l(x; L)\right].$$

Using this function, we can define the group treatment effect, similar to the individual treatment effect as,

$$T(x; L, \pi_b, \pi_e) = v(x; L, \pi_e) - v(x; L, \pi_b). \tag{2}$$

Note that the group treatment effect is constant within every l_i, and we refer to each partition l_i as a group. With little abuse of notation we denote the individual treatment effect by $t(x)$ and group treatment effect by $T(x; L)$ and may interchangeably use group and subgroup.

4.1. Group treatment effect estimator

Given a partition L, a set of trajectories $\mathcal{T}$, the behaviour policy π_b and an evaluation policy π_e the following estimator defines the group treatment effect estimator for an initial state x over a dataset $\mathcal{D} = \{(x_0, \rho_0, g_0), \ldots, (x_N, \rho_N, g_N)\}$,

$$\hat{T}(x; L) = \frac{1}{|\{x_i | x_i \in l(x; L)\}|} \sum_{i | x_i \in l(x; L)} (\rho_i g_i - g_i) \tag{3}$$

Where, $x_i = x_0^i$ is the initial state of a trajectory τ_i, g_i is the discounted return $g_i = \sum_{t=0}^{H} \gamma^t r_t^i$ and ρ_i is the importance sampling ratio $\rho_i = \prod_{t=0}^{H} \frac{\pi_e(a_t^i | x_t^i)}{\pi_b(a_t^i | x_t^i)}$. It is straightforward to show that $\hat{T}(x; L)$ is an unbiased estimator of $T(x; L)$ in every group.

Following much of the literature (Athey and Imbens, 2016; Thomas and Brunskill, 2016) we focus on the MSE criteria to rank different estimators; however, as explained later, we modify this loss in multiple ways.

$$MSE(\hat{T}; L) = \mathop{\mathbb{E}}_{x \sim \mathcal{X}} \left[\Big(t(x) - \hat{T}(x; L)\Big)^2\right] \tag{4}$$

Note that MSE loss is infeasible to compute, as we do not observe treatment effect $t(x)$. However, we show that it is equivalent to an expectation over quantities that can be estimated from data.

Theorem 1 *For a given partition $L \in \Pi$, let $T(x; L)$ be the group treatment effect defined in equation 2,*

$t(x)$ be the individual treatment effect as defined in equation 1 and $\hat{T}(x;L)$ an unbiased estimator of $T(x;L)$. The following equation imposes the same ranking over the partitions as the MSE loss in equation 4:

$$-\mathbb{E}_{x\sim\mathcal{X}}\left[\hat{T}^2(x;L)\right] + 2\ \mathbb{E}_{x\sim\mathcal{X}}\left[\mathbb{V}\left[\hat{T}(x;L)\right]\right] \quad (5)$$

Where $\mathbb{V}[\hat{T}(x;L)]$ is the variance of the estimator $\hat{T}(x;L)$.

The proof is provided in the supplementary materials.

The result of theorem 1 suggests an estimable quantity that can be used to select between different potential partitions. More precisely, given a dataset $\mathcal{D}$, the empirical version of the adjusted MSE in Equation 5 can be written as

$$EMSE(\hat{T};L) = -\frac{1}{N}\sum_{i=1}^{N}\hat{T}^2(x_i;L) + \frac{2}{N}\sum_{i=1}^{N}\mathbb{V}\left[\hat{T}(x_i;L)\right], \quad (6)$$

where $\mathbb{V}[\hat{T}(x_i;L)]$ is the variance of the estimator $\hat{T}(x_i;L)$ in the subgroup l_i s.t. $l_i = l(x_i;L)$.

5. A Practical and Effective Algorithm for Subgroup Identification

In this section, we first assume access to a loss function $\mathcal{L}(L)$ and describe the recursive partitioning algorithm to minimize it. We then discuss the modifications we apply to the empirical adjusted MSE in section 5.2 to obtain the loss function $\mathcal{L}(L)$.

5.1. Algorithm

In order to partition the feature space to different subgroups we minimize a loss function $\mathcal{L}(L)$ with recursive partitioning, $\min_{L\in\Pi}\ \mathcal{L}(L)$. First, in the partitioning phase, similar to classification and regression tree (CART) (Breiman et al., 1984), we build a tree by greedily splitting the feature space to minimize the loss function. We stop splitting further when there is no such split that results in a reduction of the loss function (partitioning phase), we call this a treatment effect tree.

After building the treatment effect tree, each leaf l_i is a group and we can form an estimate of the group treatment effect using the importance-sampling like estimator specified in Equation 3 (estimation phase). In this work, we use the same estimator in the partitioning and estimation phase and mainly focus on developing a loss function to be used in the partitioning phase. Additionally, we compute confidence intervals around our estimation by bootstrapping. Note that in the estimation phase, instead of the importance sampling approach we use here, it would also be possible to substitute different off-policy evaluation methods, such as model-based and doubly robust (Thomas and Brunskill, 2016; Liu et al., 2018) to estimate the treatment effects.

5.2. Loss Function

One way to estimate the empirical adjusted MSE in equation 6 is by substituting the variance term with the sample variance of the estimator. This is similar to the loss proposed by Athey and Imbens (2016) in the non-sequential setting. However, estimation of the sample variance may be very noisy due to limited data, particularly in our sequential setting. Misestimation of the variance may result in an avoidable undesirable split in the partitioning phase that would have not happened given a better estimate of the variance. Indeed, over-splitting is a common failure mode of using this loss function as we demonstrate in our experiments.

Variance Estimation. To mitigate the issue of over-splitting, we modify the loss function by a proxy of the variance term that can be used in limited data settings and can be efficiently computed.

First, we show that the variance of the treatment effect estimator can be upper bounded by quantities that one can compute from data.

Theorem 2 *Given a dataset $\mathcal{D} = \{(x_0,\rho_0,g_0),\dots,(x_N,\rho_N,g_N)\}$ and the treatment effect estimator defined by $\hat{T} = \frac{1}{N}\sum_i(\rho_i - 1)g_i$. The variance of $\hat{T}$ satisfies the following inequality,*

$$\mathbb{V}[\hat{T}] \leq \|g\|_\infty^2\left(\frac{1}{ESS} - \frac{1}{N}\right) \quad (7)$$

where, ESS is the effective sample size.

Note that in the special case of behaviour policy being the same as the evaluation policy, this bound evaluates to zero. We denote the RHS of equation 7 by $\mathbb{V}_u[.]$. In our work, we use $\mathbb{V}_u[.]$ in each leaf as a proxy of variance of the estimator in the leaf. That is,

$$\mathbb{V}_u[\hat{T}(x_i;L)] = ||g(x)||_\infty^2\left(\frac{1}{ESS(l_i)} - \frac{1}{|l_i|}\right), \quad (8)$$

where we use the common ESS estimate $ESS(l_i)$ by $\widehat{ESS}(l_i) = \frac{(\sum_j \rho_j)^2}{\sum_j \rho_j^2}$ where the sum is over samples inside the group i, $\{j|x_j \in l_i\}$ (Owen, 2013). $\mathbb{V}_u[.]$ can be computed efficiently and the conservative variance estimation using $\mathbb{V}_u[.]$ avoids the problem of variance underestimation.

Regularization In many applications, actionable information needs to satisfy certain conditions. For example, a clinician may consider the knowledge of group treatment effect useful, if we can guarantee with high probability that the treatment effect is α bounded away from zero. The loss function which is focused on minimizing the mean squared error would not necessarily identify these practically relevant subgroups.

Therefore we now introduce a regularization term into our loss function to encourage finding such partitions where some subgroups have treatment effects that are bounded away from zero. To do so we use Cantelli's inequality to derive a lower bound on the estimator defined in equation 3. While this is a weaker bound than Bernstein, this allows us to avoid assuming we have access to an upper bound on the importance weights.

We assume that the function $\hat{T}[x; L] : \mathcal{X} \to \mathbb{R}$, is a bounded function ($||\hat{T}(x; L)||_\infty < \infty$). We start by writing Cantelli's inequality applied to the random variable $\hat{T}(x; L)$,$\mathbb{P}\left(\hat{T}(x; L) - \mathbb{E}[\hat{T}(x; L)] \geq \lambda\right) \leq \frac{1}{1+\frac{\lambda^2}{\mathbb{V}[\hat{T}(x;L)]}}$. Assigning δ to the right hand side and considering the complementary event, we have with probability $1 - \delta$,

$$\mathbb{E}[\hat{T}(x; L)] \geq \hat{T}(x; L) - \sqrt{\frac{1-\delta}{\delta}\mathbb{V}[\hat{T}(x; L)]} \tag{9}$$

With Equation 9 we define the (margin α) regularization term

$$\mathcal{R}(x_i; L, \alpha) = \tag{10}$$
$$\max\left\{0, \alpha - \left(|\hat{T}(x_i; L)| - c\sqrt{\mathbb{V}_u[\hat{T}(x_i; L)]}\right)\right\},$$

Where c adjusts the penalty for having more limited data and provides a more robust lower bound on the estimated treatment effect, and α allows medical professionals to specify a minimum threshold on a meaningful treatment effect. Note that we used $\mathbb{V}_u[.]$ instead of $\mathbb{V}[.]$ in equation 10 to avoid issues arising from under estimation of the variance. Although we can obtain c by setting a specific value of δ, we view this as a tuning parameter for regularization.

Loss Function By combining the regularization term and using the proxy variance, we obtain our final loss function.

$$\mathcal{L}(L) = -\frac{1}{N}\sum_{i=1}^{N}\hat{T}^2(x_i; L) + \frac{2}{N}\sum_{i=1}^{N}\mathbb{V}_u\left[\hat{T}(x_i; L)\right] + \frac{C}{N}\sum_{i=1}^{N}\mathcal{R}(x_i; L, \alpha), \tag{11}$$

where C is the regularization constant. This loss is minimized using recursive partitioning. We call our algorithm GIOPE, group identification in off-policy policy evaluation. Note in Theorem 1 we relied on $\hat{T}(x; L)$ be an unbiased estimate of $T(x; L)$. To accomplish this with our chosen estimator for $T(x; l)$ we use an independent set of samples for the partitioning phase and the estimation phase. The importance of sample splitting to avoid overfitting during off-policy estimation is well studied (e.g. (Craig et al., 2020; Athey and Imbens, 2016)).

6. Experiments

We illustrate how our approach allows us to partition the feature space into subgroups such that we can make confident and accurate predictions of the group treatment effect. We empirically evaluate our method in sequential decision-making settings, compare it to the baseline and perform ablation analysis to show the benefit of each modification we have proposed.

We evaluate our method on a sepsis management simulation (Oberst and Sontag, 2019). Additionally, we use the publically available MIMIC III dataset of ICU patients (Johnson et al., 2016) and focus on the sepsis cohort (Komorowski et al., 2018) to show how our method can be applied to real-world data. We provide the code for all experiments in the supplementary materials. We compare to causal forests (CF) (Athey et al., 2019) that were developed for non-sequential settings. To the best of our knowledge, CF is one of the best performing algorithms in non-sequential settings that yields a good performance across different domains.

6.1. Sepsis Simulation

A growing number of works seek to learn an automated policy to manage septic patients in the ICU. The reader may find a short review in Gottesman et al. (2019). However, newly suggested decision policies may be beneficial for some subgroups of patients

while harmful to others. We use the sepsis simulator developed in Oberst and Sontag (2019) to demonstrate this scenario and evaluate our models in detecting such subgroups.

Simulator In this simulator, each patient is described by four vital signs {heart rate, blood pressure, oxygen concentration, glucose level} that take values in a subset of {very high, high, normal, low, very low}. There is also a binary indicator of diabetes. that results in a state space of size $|S| = 1440$. In each step, the agent can take an action to put the patient on or off of treatment options, {antibiotics, vasopressors, and mechanical ventilation}. Each episode terminates upon death with reward -1, or discharge with reward $+1$ or runs H steps with reward 0. We use a discount factor of $\gamma = 0.99$ and average over 15 different runs.

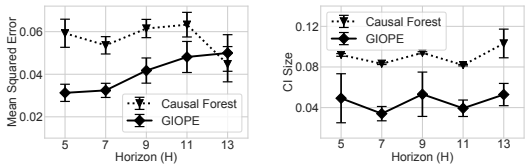
Data Generation We design the behavior and the evaluation policy to be similar and nearly optimal. More precisely, define π_{st} as the (deterministic) optimal policy for this environment, softened by subtracting 0.1 probability from the optimal action and equally distributing it among other actions.

The behavior policy π_b is similar to π_{st} except it has 15% less chance of using the mechanical ventilator. The evaluation policy π_e is similar to π_{st} but has 20% less chance of using the vasopressor: therefore it uses the mechanical ventilator more and vasopressor less than the behaviour policy. Regardless of the horizon, the evaluation policy achieves a better expected discounted return than the behaviour policy. However, there are subgroups of individuals, for example, diabetics, that will be worse off by using the evaluation policy. We generate $50,000$ trajectories using the behavior policy.

Comparison We compute the mean squared error for individual treatment effects. To do so, for each individual in the test set that consists of $n = 20,000$ samples from the same distribution as the training set, we sample 30 different trajectories using the evaluation policy and the behaviour policy to compute the true treatment effect for each individual.

Figure 1 shows the mean squared error of prediction made by our method versus the causal forest (CF) method. As shown in Figure 1 (a), our method outperforms the baseline but as the horizon increases, both models struggle to generate valid results.

Panel (b) of Figure 1 shows the average size of the 95% confidence intervals, which are substantially



(a) (b)

Figure 1: **Sepsis simulator:** comparison with causal forest (CF). (a) Mean squared error of prediction. (b) The average size of the 95% confidence intervals (CI).

tighter than the baseline causal forest method. This highlights one of the main benefits of our method: it yields more accurate predictions along with tighter confidence intervals.

Identified Subgroups As hoped, our method can reliably identify subgroups with some significant treatment effects. For example, as stated above, diabetics have a negative treatment effect and our method uncovers this subgroup. The ability of our method to lend itself to such qualitative analysis is a big advantage over other algorithms such as causal forests, as they are not designed to yield distinctive subgroups.

Ablation Study To showcase the benefit of each modification that we proposed, we perform ablation study on the sepsis simulator. We compare three different methods with 15 different runs.

1. *GIOPE*: Using the loss function presented in equation 11. In all experiments, the value of regularization is set to $C = 5.0$ and the margin $\alpha = 0.05$. We found that changing this regularization value has little effect on the results.

2. *GIOPE - Regularization (GIOPE-R)*: Using the loss function in equation 6 with the suggested proxy variance in equation 8.

3. *GIOPE - Regularization and Proxy Variance (GIOPE-RP)*: This method uses the loss function presented in equation 6 with the sample variance estimate. Note that this basic version is similar to the loss function proposed in Athey and Imbens (2016).

First, we look at the mean squared error computed on the individual level. As shown in Figure 2

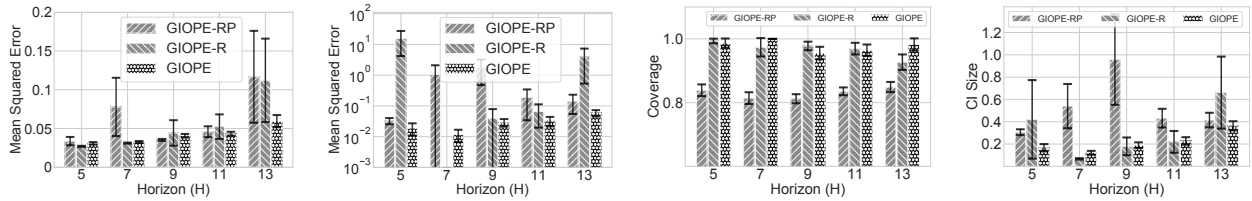


Figure 2: Ablation study. (a) Mean squared error computed on the individual level. (b) Group mean squared error. Coverage: (c) Percentage of groups that the true group treatment effect is covered by the 95% confidence interval. (d) The average size of confidence intervals

(a) our method shows significant benefits compared to GIOPE-R and GIOPE-RP. This comes with an important observation that our method also shows more stability as the performance does not fluctuate as much across different horizons as well as having smaller standard errors. Note that, our method does not optimize for this objective and the individual mean squared error is best minimized with the sample variance in the limit of infinite data, the benefit comes as with avoiding predicting each individual separately.

Next, we look at the mean squared error in the group treatment effect (Figure 2 (b)). That is, for a groups i, denote the prediction of the group treatment effect by $\hat{g}_i$ and the true group treatment effect by g_i, then the group MSE is defined as $\frac{1}{G}\sum_{i=1}^{G}(g_i - \hat{g}_i)^2$, where G is the total number of groups. Figure 2 (b) shows the MSE in group treatment effect as we increase the horizon. Similar to individual MSE, our method obtains lower MSE and displays more stability across different horizons. This stability is mainly due to avoiding oversplit. For example, the average number of discovered groups in the GIOPE-RP method for horizon 13 is 26 whereas for other GIOPE-R is 5 and GIOPE is 4.

Finally, we look at coverage. Figure 2, panel (c) shows the coverage of 95% confidence intervals of the true group treatment effect for different methods and horizons. Methods that use variance proxy instead of sample variance consistently show more coverage. Figure 2 panel (d) shows the average size of the confidence interval for each group treatment effect prediction. This indicates that using the upper bound along with regularization (GIOPE) yields more coverage while offering tighter confidence intervals. This observation highlights the main benefit of using regularization along with proxy variance that allows us to discover groups that we can more accurately and confidently predict their treatment effect. Tighter standard error of confidence intervals size highlights the stability of GIOPE across different runs.

Additionally, Figure 3 shows the partitions found in horizon $\{5, 9, 11, 13\}$. Our method can recover groups with significant negative treatment effect in different horizons.

Effect of hyper-parameters We used the following set of hyper-parameters for the experiments presented earlier. Regularization constant $C = 5.0$, regularization margin $\alpha = 0.05$, regularization confidence value $\delta = 0.4$, maximum depth of the tree $d = \infty$ and minimum number of samples in each leaf 50.

In order to evaluate the effect of hyper-parameters, we perform the ablation study for two different values of regularization confidence interval $\delta = 0.1$ and $\delta = 0.5$ and two different values of regularization constant $C = 5.0$ and $C = 1.0$. Figure 4 (a) shows the result for mean squared error and (b) for group mean squared error. As shown, the effect of regularization is small, and the same results can be obtained with a different range of hyper-parameters. Similarly, figure 4 (c) shows the coverage of the 95% confidence interval and (d) is the average size of CI. The results obtained previously hold with different hyper-parameters.

6.2. ICU data - MIMIC III

To show how our method can be used on a real data set, we use a cohort of septic patients in the freely accessible MIMIC III dataset (Johnson et al., 2016). Prior work Komorowski et al. (2018) used off-policy

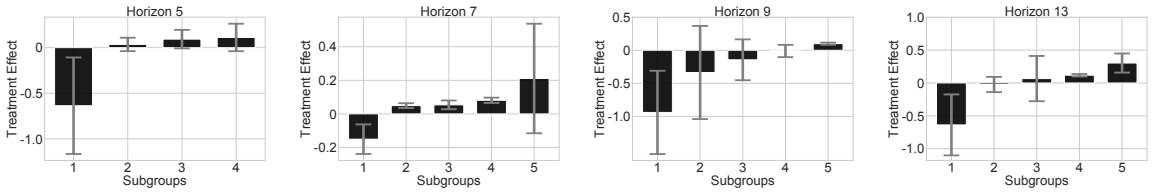


Figure 3: Sepsis simulation. Group treatment effect for a sample run with horizon (a) 5, (b) 7, (c) 9 (d) 13

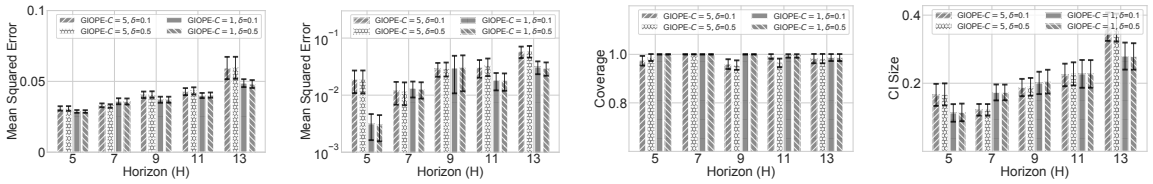


Figure 4: Effect of hyper-parameters: Ablation study, results of GIOPE for four different values of parameters. (a) Mean squared error (b) group mean squared error, (c) 95% confidence interval coverage and (d) average size of confidence intervals

learning and proposed a new decision policy that might provide improved patient outcomes on average. We followed Komorowski et al. (2018) to extract the sepsis cohort. Our training set consists of 14971 individuals, with 8442 males and 6529 females. The mortality rate in our cohort is 18.4%. The feature space is of size 44 and consists of the following values:

{`gender`, `re_admission`, `mechvent`, `age`, `Weight_kg`, `GCS`, `HR`, `SysBP`, `MeanBP`, `DiaBP`, `RR`, `Temp_C`, `FiO2_1`, `Potassium`, `Sodium`, `Chloride`, `Glucose`, `Magnesium`, `Calcium`, `Hb`, `WBC_count`, `Platelets_count`, `PTT`, `PT`, `Arterial_pH`, `paO2`, `paCO2`, `Arterial_BE`, `Arterial_lactate`, `HCO3`, `Shock_Index`, `Shock_Index`, `PaO2_FiO2`, `cumulated_balance`, `SOFA`, `SIRS`, `SpO2`, `BUN`, `Creatinine`, `SGOT`, `SGPT`, `Total_bili`, `INR`, `output_total`, `output_4hourly`}

We provide the index of the patients in the dataset to facilitate the reproducibility of our results.

To estimate the behavior policy, we use KNN with $k = 100$ on the test set, we use l_2 distance with uniform weights across different features to measure the distance. If an action was not taken among all 100 nearest neighbors, we assign the probability 0.01 to the action. We used IV fluid and mechanical ventilation for actions and used 20% quantile to discretize the action space into 25 actions.

For the evaluation policy, we used a similar method as the behavior policy on a random subset of the training set (20% of the training data). We only used the following features to estimate the distance for the evaluation policy,

{`HR`, `SysBP`, `Temp_C`, `Sodium`, `Chloride`, `Glucose`, `Calcium`, `paO2`, `Arterial_BE`, `SOFA`, `SIRS`, `Creatinine`}

Similarly, if an action was not taken among all 100 nearest neighbors, we assign the probability 0.01 to the action. In our experiments we used the following set of hyper-parameters: regularization constant $C = 100.0$, regularization margin $\alpha = 0.0$, regularization confidence value $c = 2.0$, maximum depth of the tree $d = \infty$ and minimum number of samples in each leaf 1000.

Using weighted importance sampling the estimated value of the decision policy is 65.33 with the effective sample size of 146.8 which suggests an increase of 2.43 on the survival chance compared to the behavior policy. Here we take this decision policy and estimate its impact on different potential subgroups.

In Figure 5 (a) we present the five groups produced by our algorithm along with their estimated

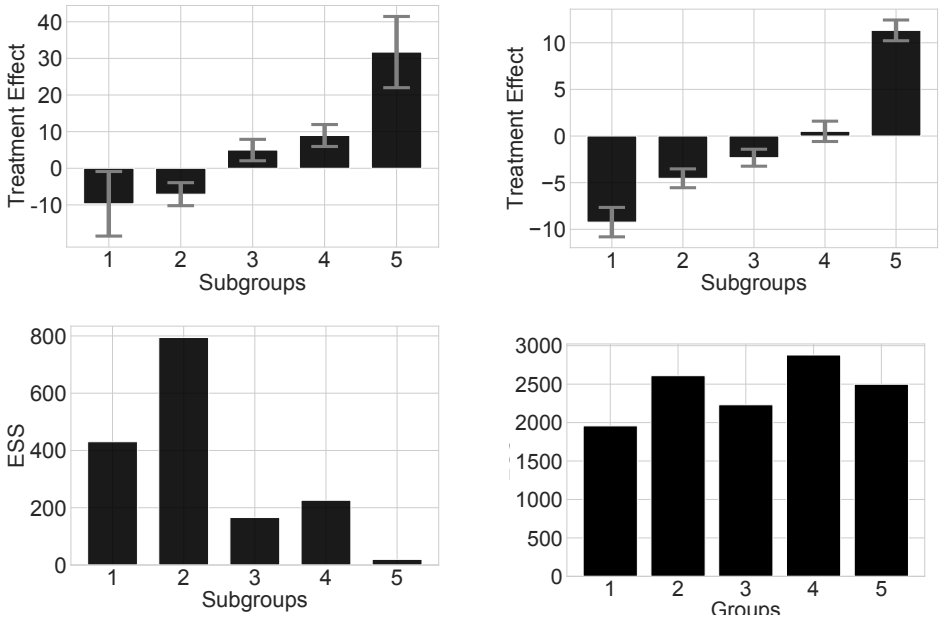


Figure 5: MIMIC III dataset. AI Clinician: Although positive treatment effect is predicted by weighted importance sampling on the full cohort, groups 1 and 2 will likely be harmed by the evaluation policy. (top) : The estimated treatment effect for each subgroup, (bottom): the estimated effective sample size (bottom). (a): Evaluation policy, (b): Evaluation policy with more aggressive vasopressor use.

group treatment effect (which is the difference between the baseline clinician policy and the decision policy). While some of the patients fall into subgroups 3, 4, and 5, some patients may experience no benefit or even a potential negative treatment effect from the proposed new treatment policy (groups 1 and 2). This highlights how our method may be useful in identifying subgroups in which a new decision policy may be beneficial or harmful relative to the standard approach. Please refer to the supplementary materials for information about the effective sample size (ESS) of each group.

We caveat the results in this section by noting that using IS based methods on real-world datasets, and the MIMIC III dataset, in particular, is very susceptible to noise-induced by the small effective sample size of the cohort (Gottesman et al., 2018). Furthermore, our method is susceptible to this source of noise twice, as IS based estimators are used both in the partitioning phase and the estimation phase. However, despite their high susceptibility to noise, IS methods are often applied to the MIMIC III dataset for their theoretical properties, but their results for real data should be interpreted with caution. In our experiment, we intentionally designed the decision policy close to the behavior policy to avoid issues arising from a small effective sample size.

More Aggressive Use of Vasopressor Additionally, we evaluated our method on a policy that utilizes vasopressor more often than the estimated behavior policy. We estimate the behaviour policy using KNN, and evaluate a policy that has 10% more probability mass on using vasopressor than the behaviour policy. Figure 5 (b) shows the subgroups found using our method.

Group 1, 2, and 3 all show a negative treatment effect. Interestingly, these three groups have `SOFA` > 1 which indicates these patients are at high risk. Given the discussions we had with an intensivist, this is in agreement with their expectation that healthier patients are less likely to be harmed by more aggressive use of vasopressor, and sicker patients may be more

at risk. This also highlights one of the main benefits of our method: it can be used to provide interpretable subgroups with different potential treatment effects that may be used to support communication with clinicians around potentially beneficial alternate treatments, and who they might benefit from.

7. Conclusion

In this paper, we proposed a novel method to partition the feature space, enabling us to find subgroups that we can accurately and confidently predict the group treatment effect for them. Our approach is in contrast with previous methods that estimate individual-level treatment effects, yielding uncertain and less accurate predictions. We do so, by proposing a novel loss function that utilizes; 1. A proxy on the variance estimator that is easy to compute and stable; 2. A regularization term that incentivizes the discovery of groups with treatment effect sizes that are considered to be significant, which may be specified by a domain expert. We further evaluate our method on both simulated domains and real-world data.

Our method can leverage the existing data to raise caution when necessary about a possible negative effect of the newly suggested decision policy on some subgroups. While it may suggest also the potential positive effect of certain policies on certain subgroups, such findings should be validated via a randomized clinical trial or form just one part of the array of information used by a medical team to inform treatment.

A particularly promising potential benefit of our approach is that results from our method, when applied to observational data, can help to design multistage randomized trials that are powered toward detecting harm or benefit of the evaluation policy compared to the baseline policy for specific subgroups.

Institutional Review Board (IRB)

This research do not require IRB approval.

Acknowledgments

The research reported here was supported in part by DEVCOM Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196, and NSF IIS-2007076.

References

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

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer, 2010.

Erin Craig, Donald A Redelmeier, and Robert J Tibshirani. Finding and assessing treatment effect sweet spots in clinical trial data. *arXiv preprint arXiv:2011.10157*, 2020.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 243–263, 2014.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Augustine Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.

Hyun-Suk Lee, Yao Zhang, William Zame, Cong Shen, Jang-Won Lee, and Mihaela van der Schaar. Robust recursive partitioning for heterogeneous treatment effects with uncertainty quantification. *Advances in Neural Information Processing Systems*, 33:2282–2292, 2020.

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation. *arXiv preprint arXiv:1805.09044*, 2018.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. *NeurIPS*, 2018.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

Appendix A. Proofs

Here we present the proof of the theorem 1. We restate the theorem first.

Theorem *[1] For a given partition $L \in \Pi$, let $T(x;L)$ be the group treatment effect defined in equation 2, $t(x)$ be the individual treatment effect as defined in equation 1 and $\hat{T}(x;L)$ an unbiased estimator of $T(x;L)$. The following equation imposes the same ranking over the partitions as the MSE loss in equation 4:*

$$-\mathbb{E}_{x\sim\mathcal{X}}\left[\hat{T}^2(x;L)\right] + 2\ \mathbb{E}_{x\sim\mathcal{X}}\left[\mathbb{V}\left[\hat{T}(x;L)\right]\right] \quad (12)$$

Where $\mathbb{V}[\hat{T}(x;L)]$ is the variance of the estimator $\hat{T}(x;L)$.

Proof First, We form the adjusted MSE (AMSE) as

$$AMSE(\hat{T};L) = \mathop{\mathbb{E}}_{x\sim\mathcal{X}}\left[\left(t(x) - \hat{T}(x;L)\right)^2 - t(x)^2\right]$$

Adjusted MSE and MSE impose the same ranking among different partitioning as $\mathop{\mathbb{E}}_{x\sim\mathcal{X}}[t(x)^2]$ is independent from the partitioning. Note that adjusted MSE, similar to MSE cannot be computed.

We continue by decomposing the adjusted MSE by adding and subtracting $T(x;L)$,

$$\begin{aligned}AMSE(\hat{T};L) &= -\mathbb{E}_{x\sim\mathcal{X}}\left[\left(t(x)-\hat{T}(x;L)\right)^2 - t(x)^2\right] \\ &= \mathbb{E}_{x\sim\mathcal{X}}[\underbrace{(t(x)-T(x;L))^2 - t(x)^2}_{(i)} \\ &+ \underbrace{(T(x;L)-\hat{T}(x;L))^2}_{ii} \\ &+ \underbrace{2(t(x)-T(x;L))(T(x;L)-\hat{T}(x;L))}_{(iii)}]\end{aligned}$$

Now we look at each part separately, for part (i),

$$\begin{aligned}&\mathbb{E}_{x\sim\mathcal{X}}\left[(t(x)-T(x;L))^2 - t(x)^2\right] = \\ &\mathbb{E}_{x\sim\mathcal{X}}\left[T(x;L)^2 - 2t(x)T(x;L)\right]\end{aligned}$$

Now we expand the expectation over each group of thr partition $L=\{l_1,\ldots,l_M\}$,

$$\sum_{l_i\in L} P(l_i)T(x;l_i)^2 - 2\sum_{l_i\in L} P(l_i)T(x;l_i)$$

Where $T(x;l_i)=T(x;L)$ such that $x\in l_i$, note that by definition, $T(x;l_i)$ is constant for all $x\in l_i$. Next, note that $\mathbb{E}_{x\in l_i}[t(x)]=T(x;l_i)$.

$$\begin{aligned}&\sum_{l_i\in L} P(l)T(x;l_i)^2 - 2\sum_{l_i\in L} P(l_i)T(x;l_i)^2 = \\ &-\sum_{l_i\in L} P(l_i)T(x;l_i)^2 = -\mathbb{E}_{x\sim\mathbb{X}}[T(x;L)^2] \qquad (13)\end{aligned}$$

Now, consider the variance of $\hat{T}(x;l_i)$ for group $l_i\in L$,

$$\begin{aligned}\mathbb{V}\left[\hat{T}(x;l)\right] &= \mathop{\mathbb{E}}_{x\in l_i}[\hat{T}^2(x;l_i)] - \left[\mathop{\mathbb{E}}_{x\in l_i}\hat{T}(x;l_i)\right]^2 \\ &= \mathop{\mathbb{E}}_{x\in l_i}[\hat{T}^2(x;l_i)] - T(x;l_i)^2 \qquad (14)\end{aligned}$$

Which follows by $\hat{T}(x;l_i)$ being an unbiased estimator of $T(x;l_i)$. Taking the expectation over the feature space and substituting equation 14 into 13,

$$\begin{aligned}(i) &= -\sum_{l_i\in L} P(l_i)T(x;l_i)^2 \\ &= \sum_{l_i\in L} P(l_i)\left[\mathbb{V}\left[\hat{T}(x;l_i)\right] - \mathbb{E}_{x\sim l_i}\left[\hat{T}^2(x;l_i)\right]\right] \\ &= \mathbb{E}_{x\sim\mathbb{X}}\left[\mathbb{V}\left[\hat{T}(x;l)\right]\right] - \mathbb{E}_{x\sim\mathbb{X}}\left[\hat{T}^2(x;l)\right]\end{aligned}$$

Now we consider part (ii),

$$\begin{aligned}&\mathbb{E}_{x\sim\mathbb{X}}\left[\left(T(x;L)-\hat{T}(x;L)\right)^2\right] \\ &= \sum_{l_i} P(l_i)\mathbb{E}_{x\in l_i}\left[\left(T(x;l_i)-\hat{T}(x;l_i)\right)^2\right] \\ &= \sum_{l_i} P(l_i)\mathbb{E}_{x\in l_i}\left[\left(\mathbb{E}_{x\in l_i}[\hat{T}(x;l_i)]-\hat{T}(x;l_i)\right)^2\right] \\ &= \sum_{l_i} P(l_i)\mathbb{V}\left[\hat{T}(x;l_i)\right] \\ &= \mathbb{E}_{x\sim\mathbb{X}}\left[\mathbb{V}\left[\hat{T}(x;L)\right]\right]\end{aligned}$$

Where the third line follows by $\hat{T}(x;l_i)$ being an unbiased estimator of $T(x;l_i)$.

Looking at the last term (iii),

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{X}}\left[\left(t(x)\hat{T}(x;L)\right)\right] &= \sum_{l_i\in L} P(l_i)\mathbb{E}_{x\in l_i}\left[t(x)\hat{T}(x;l_i)\right] \\ &= \sum_{l_i\in L} P(l_i)\hat{T}(x;l_i)\mathbb{E}_{x\in l_i}[t(x)] \\ &= \sum_{l_i\in L} P(l_i)\hat{T}(x;l_i)T(x;l_i) \\ &= \mathbb{E}_{x\in\mathcal{X}}\left[\hat{T}(x;l_i)T(x;l_i)\right]\end{aligned}$$

Which implies $\mathbb{E}_{x\sim\mathcal{X}}\left[(t(x)-T(x;L))\hat{T}(x;L)\right]=0$. As a result,

$$\begin{aligned}(iii) &= 2\mathbb{E}_{x\in\mathcal{X}}\left[(t(x)-T(x;L))\left(T(x;L)-\hat{T}(x;L)\right)\right] \\ &= 2\mathbb{E}_{x\in\mathcal{X}}\left[(t(x)-T(x;L))\,(T(x;L))\right] \\ &- 2\mathbb{E}_{x\in\mathcal{X}}\left[(t(x)-T(x;L))\left(\hat{T}(x;L)\right)\right] = 0\end{aligned}$$

Putting the results together, results in equation 11. ■

Next, we continue with the proof of Theorem 2. First, we present a reminder on Rényi divergence.

Rényi Divergence For $\alpha\geq 0$ the Rényi divergence for two distribution P and Q as defined by (Cortes et al., 2010) is

$$D_\alpha(P||Q) = \frac{1}{\alpha-1}\log_2\sum_x Q(x)\left(\frac{P(x)}{Q(x)}\right)^{\alpha-1}.$$

Denote the exponential in base 2 by $d_\alpha(P_e||P_b) = 2^{D_\alpha(P_e||P_b)}$.

The effective sample size (ESS) (Kong, 1992) is often used for diagnosis of IS estimators, and is defined as

$$ESS(P||Q) = \frac{N}{1 + \mathbb{V}_{x\sim Q}[w(x)]} = \frac{N}{d_2(P||Q)}$$

where N is the number of samples drawn to estimate the importance weights. A common estimator of the ESS (Owen, 2013) is

$$\widehat{ESS}(P||Q) = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}$$

Theorem *[2] Given a dataset* $\mathcal{D} = \{(x_0, \rho_0, g_0), \ldots, (x_N, \rho_N, g_N)\}$ *and the treatment effect estimator defined by* $\hat{T} = \frac{1}{N}\sum_i(\rho_i - 1)g_i$. *The variance of* $\hat{T}$ *satisfies the following inequality,*

$$\mathbb{V}[\hat{T}] \leq \|g\|_\infty^2 \left(\frac{1}{ESS} - \frac{1}{N}\right) \tag{15}$$

where, ESS is the effective sample size.

Proof First note that the variance of the treatment effect estimator $\hat{T} = \frac{1}{N}\sum_i(\rho_i - 1)g_i$ can be upper bounded by the variance of the importance sampling weights. Since $\mathbb{V}[\hat{T}] \leq \mathbb{E}[\hat{T}^2]$

$$\mathbb{V}[\hat{T}] \leq \frac{||g||_\infty^2}{N^2}\mathbb{E}\left[\sum_i(\rho_i - 1)^2\right] = \frac{1}{N}||g||_\infty^2\mathbb{V}[\rho],$$

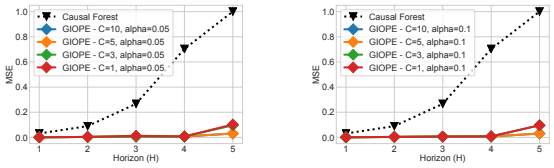
where the last equality follows by observing that $\mathbb{E}[\rho] = 1$. As noted by Metelli et al. (2018), The variance of the treatment effect estimator can be written as

$$\mathbb{V}[\hat{T}] \leq |||g||_\infty^2(\frac{d_2(P_e||P_b)}{N} - \frac{1}{N})$$

This expression can be related to the effective sample size of the original dataset given the evaluation policy, resulting in equation 15

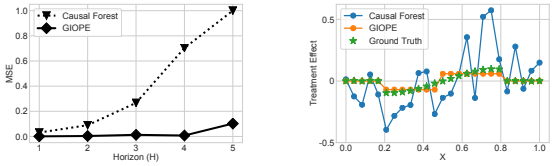
$$\mathbb{V}[\hat{T}] \leq ||g||_\infty^2 \left(\frac{1}{ESS} - \frac{1}{N}\right)$$

■



(a) $\alpha = 0.05$ (b) $\alpha = 0.1$

Figure 6: Toy MDP. (a) regularization margin $\alpha = 0.05$, (b) regularization margin $\alpha = 0.1$



(a) (b)

Figure 7: **Toy MDP:** (a) Mean squared error of treatment effect prediction for our method and causal forest(CF). (b) True and predicted treatment effect for different values of x for our method and causal forest.

Appendix B. Toy Example

In this section we present experimental results on a simple MDP.

We consider a 1 dimensional toy MDP to illustrate the difference between our method and methods developed for non-sequential setting. The toy MDP has the transition dynamics $x_{t+1} = clip(x_t + \kappa \times a_t + \epsilon, 0, 1)$, where the function $clip(x, a, b)$, clips the value of x between a and b and reward function $r(x) = 1 - |x - 0.5|$.

We consider a simple Markov decision process (MDP) with the state space $x \in [0, 1]$, discrete action space $a \in \{-1, 0, 1\}$ and the reward function is defined as $r(x) = 1 - |x - 0.5|$. The transition dynamic is specified by, $x_{t+1} = clip(x_t + \kappa \times a_t + \epsilon, 0, 1)$, where the function $clip(x, a, b)$, clips the value of x between a and b, $\kappa = 0.2$ and $\epsilon \sim \mathcal{N}(0, 0.05)$. Each episode lasts H steps. Intuitively, action 1 takes the agent to the right, -1 to the left and 0 same location with some gaussian noise. If the agent hits the boundary, the action has no effect on the position.

The behaviour policy, takes action with the following probabilities

$$\begin{cases} x < 0.2: \ \pi_b(-1) = 0.25, \ \pi_b(0) = 0.25, \ \pi_b(1) = 0.5 \\ x \geq 0.2: \ \pi_b(-1) = 0.5, \ \pi_b(0) = 0.25, \ \pi_b(1) = 0.25 \end{cases}$$

And the evaluation policy,

$$\begin{cases} x > 0.8: \ \pi_e(-1) = 0.5, \ \pi_e(0) = 0.25, \ \pi_e(1) = 0.25 \\ x \leq 0.8: \ \pi_e(-1) = 0.25, \ \pi_e(0) = 0.25, \ \pi_e(1) = 0.5 \end{cases}$$

We generated 50000 trajectories with the behaviour policy for horizons $\{1, 2, 3, 4, 5\}$ and averaged all results over 10 runs.

We look at the mean squared error of the treatment effect prediction on 25 equally spaced points in $[0, 1]$. Figure 7 (a) compares the MSE between our method with causal forest (CF). GIOPE shows smaller MSE and as the horizon increases the benefit is more apparent. Figure 7 (b) shows the predicted value of the treatment effect for our method and causal forest for horizon $H = 4$ along with the true treatment effect for different values of x. This illustrates the reason of performance gap. Our method partitions the state space and makes the same prediction for each subgroup that results in more accurate predictions, whereas causal forests over-splits and compute different values of the treatment effect for every value of x which are often inaccurate.

Figure 6 compares the mean squared error of our method versus the causal forests for different range of hyper-parameters. Panel (a) shows the results for margin $\alpha = 0.05$ and values of regularization constant $C = \{1, 3, 5, 10\}$ and panel (b) shows the results for margin $\alpha = 0.1$. As shown, regularization has small effects on the results and the results reported in the main text holds for a large range of hyper-parameters.