# MDP$^2$ Forest: A Constrained Continuous Multi-dimensional Policy Optimization Approach for Short-video Recommendation

Sizhe Yu*
School of Statistics and Management, Shanghai University of Finance and Economics
Shanghai, China
yusizhe_1@126.com

Ziyi Liu*
School of Statistics, Renmin University of China
Beijing, China
ziyiliu1125@163.com

Shixiang Wan
Tencent Inc.
Beijing, China
shixiangwan@tencent.com

Jia Zheng
Tencent Inc.
Beijing, China
zerojzheng@tencent.com

Zang Li
Tencent Inc.
Beijing, China
gavinzli@tencent.com

Fan Zhou†
School of Statistics and Management, Shanghai University of Finance and Economics
Shanghai, China
zhoufan@mail.shufe.edu.cn

## ABSTRACT

In the ecology of short video platforms, the optimal exposure proportion of each video category is crucial to guide recommendation systems and content production in a macroscopic way. Though extensive studies on recommendation systems are devoted to providing the most well-matched videos for each view request, fitting the data without considering inherent biases such as selection bias and exposure bias will result in serious issues. In this paper, we formalize the exposure proportion strategy as a policy-making problem with multi-dimensional continuous treatment under certain constraints from a causal inference point of view. We propose a novel ensemble policy learning method based on causal trees, called Maximum Difference of Preference Point Forest (MDP$^2$ Forest), which overcomes the shortcomings of existing policy learning approaches. Experimental results on both simulated and synthetic datasets show the superiority of our algorithm compared to other policy learning or causal inference methods in terms of the treatment estimation accuracy and the mean regret. Furthermore, the proposed MDP$^2$ Forest method can also adapt to a wide range of business settings such as imposing different kinds of constraints on the multi-dimensional treatment.

## CCS CONCEPTS

• **Human-centered computing** → *Social networking sites*; • **Computing methodologies** → **Causal reasoning and diagnostics**; • **Information systems** → *Recommender systems.*

## KEYWORDS

optimal policy, causal inference, forest, short-video recommender

**ACM Reference Format:**
Sizhe Yu, Ziyi Liu, Shixiang Wan, Jia Zheng, Zang Li, and Fan Zhou. 2022. MDP$^2$ Forest: A Constrained Continuous Multi-dimensional Policy Optimization Approach for Short-video Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539341

## 1 INTRODUCTION

Watching, uploading and sharing short video has become a new lifestyle all over the world. People, especially young generations, are spending more time in the endless short-video streams than ever before. The recommendation stream of short video platforms helps users to find what they are interested in, such as timing news, useful skills, and entertainment videos. Different from traditional search engines or E-commerce systems, most users of short video platforms usually explore videos without purpose. Therefore, a certain number of short-video applications use "immersive feed stream" (full-screen short-videos recommendation stream) as their main entry, which emphasize the key position of the recommendation system. However, exposing the right short-videos to a right person is challenging due to the high exposure request frequency, a large number of candidate videos, and also the sparse, noisy, implicit and imbalance offline data.

Although the empirical success achieved by existing works [10, 11, 13, 14, 16, 17, 26, 27] in improving the recommendation efficiency by generating optimal recommending video lists for any user at any particular time, they are excessively focusing on the optimal recommendations at individual level and ignore the constraint on the distributions of exposed short-videos of different categories from the global perspective. All these works pay too much attention to the micro recommendations, although their performance is evaluated using the averaged prediction accuracy. Moreover, most of these methods are vulnerable to a variety of statistical

*Both authors contributed equally to this research, and completed this work during the internship in Tencent
†Corresponding author



*KDD '22, August 14–18, 2022, Washington, DC, USA*

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539341

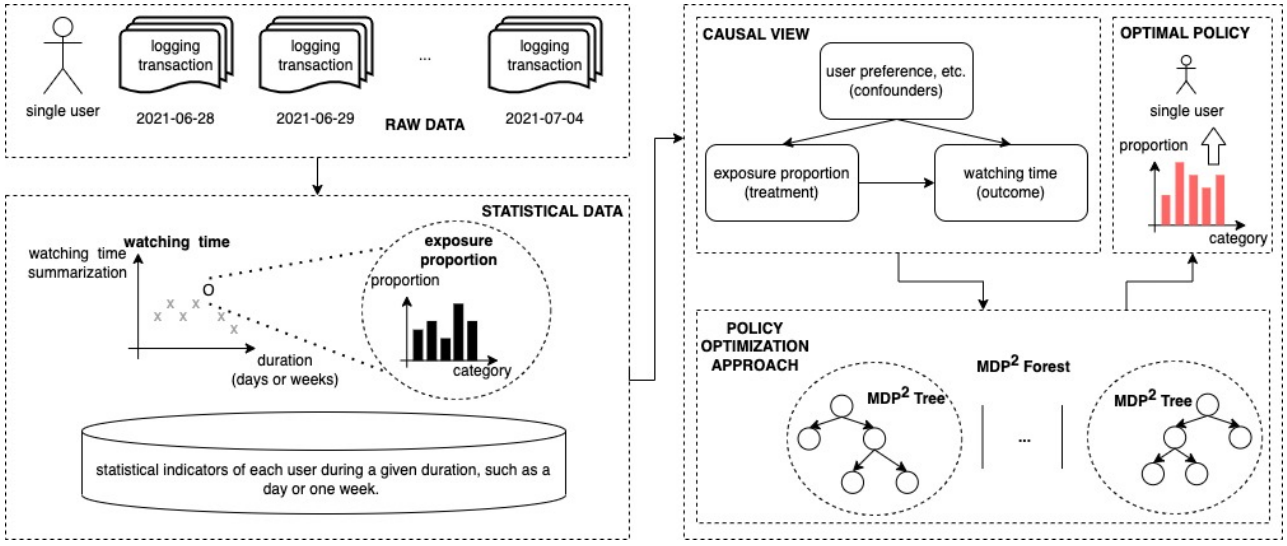



**Figure 1: The pipeline of our method. Left part is the workflow of data processing and right part is the modeling architecture. For each single user, we collect raw data from the logged transaction histories and then obtain his watching time to build the training data. Some portrait and behavioral characteristics are obtained as the user's preferences. Under the causal view, we respectively take the exposure proportions and watching time as treatment and outcome, and regard the user's preference as the confounder. Finally, we apply the proposed MDP$^2$ Forest to find the optimal policy.**

biases[6, 7, 15, 25, 29, 31, 32] when directly trained using the observed users' feedbacks from logged data. The complexity of the recommendation system may enlarge this misspecification due to ubiquitous but necessary interventions at each stage, such as rules for cold start and diversity requirements.

To reduce the interference complexity and provide an overall guidance for the recommendation requests, our method is designed in a macro view. Different from micro strategies that aim at recommending the most relevant short video to satisfy each specific request, macro strategies try to control and balance some global statistical indicators, such as the exposure proportion which represents the distribution of exposures of different short-video categories in a given time duration, as shown by Figure 1. Keeping the exposure proportions of different categories near the optimal values helps increase users' satisfactions and finally brings more platform revenue. In practice, the offline exposure proportions model is deployed at the ranking stage of recommendation, and those recommended videos should be sure not to deviate from this proportion. As a macro guidance metric, optimal exposure proportions can also benefit the content ecosystem of short-videos from two aspects. First, as an operation tool, optimal exposure proportions could make a guidance to the content production by pointing out the redundant and inadequate videos. Second, the optimal exposure proportions can help determine the optimal amount of investment on each category by the platform in a quantitative way.

Our goal is to find the personalized optimal exposure proportion and introduce a restricted recommendation policy from the macro perspective. To be specific, we treat the exposure proportion as multi-dimensional continuous treatment, whose elements are inherently correlated to each other and sum to one. Accordingly, we can formulate the short-video recommendation as a multi-dimensional policy-making problem, and one of the fundamental questions to be answered is what kind of exposure proportions (treatment) can lead to the longest watching time (outcome) for a certain user (context). The causal relationships behind are shown in the "CAUSAL VIEW" part of Figure 1. Unfortunately, very few studies simultaneously consider the exposure proportions of different video categories. The complex causal structure and the combinatorial explosion of the multi-dimensional treatments make it impossible to estimate all counterfactuals and heterogeneous treatment effects, or to choose the optimal treatment from all the candidates [8]. Moreover, the experimental results in this paper show that these methods can not perform well in short-video recommendation problems. Another related research area is offline policy learning [3, 9, 12], which aims to finding the optimal policy by evaluating target policy offline using the logged data. However, these optimization-based methods have strong assumptions about the dimension of the treatment space and is not robust to different model constraints, which impair their empirical performance. More discussion about this can be found in Section 5.

To fill this research gap, we propose MDP$^2$ Forest* , a dedicated effective tree-based policy learning algorithm based on maximum difference point of preference function that learns optimal personalized exposure proportion assignment of all categories simultaneously. We introduce a novel difference measure which generalizes the idea of [23] to discretize the continuous space. Moreover, a heuristic method combined with dimension iterations helps learn the optimal policy in the multi-dimensional treatment space with constraints ignored by previous studies. Some strategies that help reduce the complexity and increase the computation speed

*Our data and code are available at https://github.com/wheels97/MDPP_Forest

are also applied to ensure the feasibility of the algorithm on large-scale datasets. Extensive experiments on simulated, semi-synthetic data indicate that MDP$^2$ Forest significantly outperforms existing causal inference methods and other policy learning algorithms. The contributions of this paper are summarized as follows:

(i) To the best of our knowledge, we are the first to propose a macro recommendation policy for short-videos by optimizing personalized exposure proportions of video categories based on causal inference and counterfactual reasoning.
(ii) We formalize the optimization of exposure proportions as a policy learning problem with continuous multi-dimensional treatments under constraints which is seldom studied by the community. The whole framework of this work can be practically applied to different business domains, and conveniently extended to different model setups.
(iii) We propose MDP$^2$ Forest, a dedicated and effective tree-based algorithm, to learn optimal policy in the continuous multi-dimensional case, whose outer-performance over existing causal inference and policy learning methods are demonstrated by simulated and synthetic experiments.

# 2 PRELIMINARY

## 2.1 Problem Formulation and Notation

**Problem.** Suppose we have a user set $\mathcal{U} = \{u_1, u_2, \ldots, u_L\}$ and a video set $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, where each user $u_i$ can also be treated as a unit. Each video belongs to one or multiple categories $\mathbb{C}_j \subset C$, where $C = \{c_1, c_2, \ldots, c_K\}$ is the set of all categories. Detailed logging data $\mathcal{L} = \{(u_i, v_{i,j}, y_{i,j})\}_{i\in\mathcal{U}, j\in\mathcal{V}}$ can be collected from the transaction system.

Different from other recommendation algorithms which are designed from the micro perspective and focus on $\mathcal{L}$, we instead learn in a macro way using the logging data $\mathcal{D}_d = \{(\boldsymbol{x_i}, \boldsymbol{t_i}, y_i\}_{i=1}^n$ within a certain time duration $d$, where $d$ can be one day, one week, etc. For each triplet $(\boldsymbol{x_i}, \boldsymbol{t_i}, y_i)$, $\boldsymbol{x_i} \in \mathbb{R}^D$ denotes the features of $u_i$, the treatment $\boldsymbol{t_i} = (t_{i1}, \ldots, t_{iK})^\top \in \mathcal{T} \subset [0,1]^K$ contains the exposure proportion of each video category for user $u_i$ within time duration $d$, and the outcome $y_i \in \mathbb{R}$ is the total or average watching time for $u_i$. For the ease of notation, we omit the subscript $d$ in $\mathcal{D}_d$ in following sections without loss of generality.

**Target.** As shown in the "CAUSAL VIEW" part of Figure 1, $\boldsymbol{t_i}$ is treated as a multi-dimensional continuous treatment with the constraint that the $K$ exposure proportions sum to 1, i.e. $\sum_{k=1}^K t_{ik} = 1$. A personalized policy is a mapping $\pi : \mathbb{R}^D \to \mathcal{T}$, which assigns treatment $\boldsymbol{t} = \pi(\boldsymbol{x})$ to a user with feature $\boldsymbol{x}$. The logging policy is unknown since we do not participate in the historical platform operations in most cases. Our goal is to find optimal policy $\pi^* = argmax_\pi \mathbb{E}(y(\pi(\boldsymbol{x})))$, or equally, to minimize the regret.

**Generalization.** The formulation above can be extended to other business or research domains as well. Each element of the treatment $\boldsymbol{t_i}$ can even be a vector, which make $\boldsymbol{t_i}$ a tensor. This setting can be widely used in robotics, healthcare or other fields which require fine controls. In this case, the treatment is a combination of several multi-dimensional continuous treatments. The constraint can be modified to ${t_i}^T\beta \le C$ or ${t_i}^T\beta = C$ and the whole framework can be generalized to the more complicated multi-constraint case.

## 2.2 Causal Assumption

In this work, we model the short-video recommendation problem from the perspective of causal inference by making several important assumptions[28] as follows.

Assumption 2.1. ***Stable Unit Treatment Value Assumption (SUTVA).*** *Units (platform users) do not interfere with each other, and treatment levels are well defined.*

From the macro perspective, users' preferences are inherent and the optimal exposure proportion of each single user is not affected by other users, which makes SUTVA hold.

Assumption 2.2. ***Ignorability.*** *Given the user's features, $\boldsymbol{X}$, the treatment assignment $\boldsymbol{T}$ is independent of the potential outcomes, which is mathematically formalized as*

$$y(\boldsymbol{t}) \perp\!\!\!\perp \boldsymbol{T}|\boldsymbol{X},$$

This assumption ensures that the exposure proportions for each user are exogenous and there are no unmeasured confounding variables.

Assumption 2.3. ***Positivity.*** *$\forall \mathrm{t} \in \mathcal{T}$, there exists $\epsilon > 0$ s.t. the conditional density (propensity) $p(\mathrm{t}|\mathrm{x}) \ge \epsilon$ with probability 1.*

The Positivity ensures that any particular exposure proportion can be assigned to any user with a non-zero possibility.

# 3 APPROACH

In this section, we introduce a novel tree-based algorithm, Maximum Difference Point of Preference (MDP$^2$) Forest to learn the optimal recommending policy, where all trees in the forest are trained in parallel.

For most real-world policy-making problems such as personalized recommendations and precision medicine, we can reasonably assume that the whole population can be divided into several subgroups, such that users' preferences (video interests, optimal doses, etc.) are homogeneous within each subgroup and heterogeneous across different subgroups. Therefore, a good tree model can effectively divide the training data into the same subgroups as the underlying ground-truths through the hierarchical structures.

## 3.1 Splitting criterion for binary treatment allocation

We first introduce the splitting criterion for a single tree in MDP$^2$ Forest under the binary treatment setting, which will be then extended to the multi-dimensional case. Suppose there are $M$ leafs or subgroups in a tree, then we can obtain a partitioning of the training samples, denoted by $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_M\}$, where $\cup_{i=1}^M \mathcal{G}_i = \{1, 2, \cdots, n\}$ and $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$ for any $i \ne j$. [1] proposes a splitting criterion that minimizes the expected mean squared error (EMSE), $\mathbb{E}(\mathrm{Y} - \hat{\mu}(\mathrm{X}; \mathcal{D}, \mathcal{G}))^2$, where $\hat{\mu}(\mathrm{X}; \mathcal{D}, \mathcal{G}) = \frac{1}{|\mathcal{G}^{\mathrm{X}}|}\sum_{\mathrm{i}\in\mathcal{G}^{\mathrm{X}}} \mathrm{y_i}$ and $\mathcal{G}^X$ is the subgroup to which X belongs. Following [1], the splitting

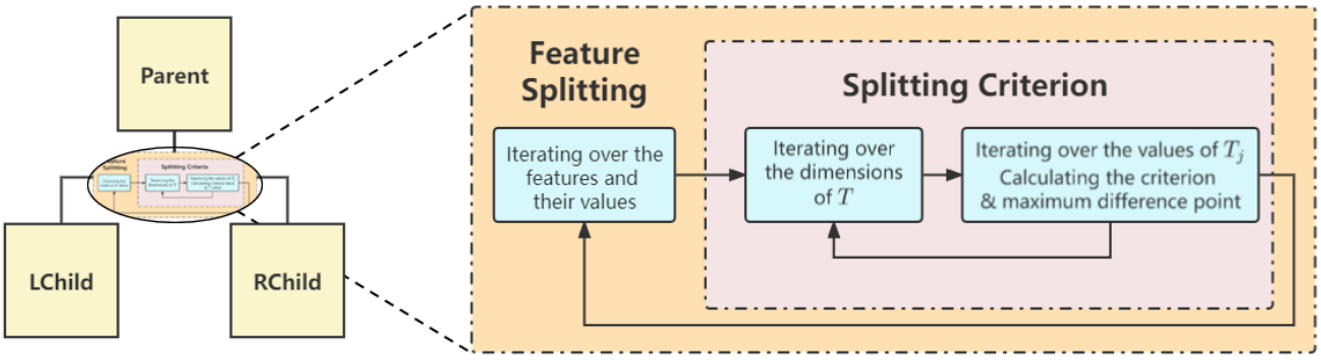



**Figure 2: Flow chart of the node splitting. Discretization is employed to calculate the node entropy (i.e. difference). For multi-dimensions $\{T_i\}_{i=1}^K$, we iterate over the dimensions and take the sum of the optimal outcomes of each dimension as the objective of the splitting criterion.**

criterion of the $i$-th partition proposed by [23] for the binary treatment is defined as follows,

$$|\mathcal{G}_i|(\hat{\mu}_1(\mathrm{X};\mathcal{D},\mathcal{G}_i)-\hat{\mu}_0(\mathrm{X};\mathcal{D},\mathcal{G}_i))^2, \tag{1}$$

where $\hat{\mu}_t(\mathrm{X};\mathcal{D},\mathcal{G}_\mathrm{i}) = \frac{1}{|\mathcal{G}_\mathrm{i}|}\sum_{\mathrm{j}\in\mathcal{G}_\mathrm{i},\mathrm{t_j}=\mathrm{t}}\mathrm{y_j}$ is the conditional mean of the $i$-th partition under $T = t \in \{0,1\}$. This criterion can not be directly applied to the multi-dimensional continuous treatment situation and the information about the optimal treatment can not be incorporated. We will describe in detail how we can address these two issues in the following subsection.

## 3.2 Optimal treatment for each subgroup

Our analysis is based on the assumption that preferences are homogeneous within each subgroup. We extract the concept of the preference as a function of the $i$-th dimension of the treatment space $\mathcal{T}_i$, $i = 1, \cdots, K$. For an observation x in the partition $\mathcal{G}_{m_\mathrm{x}}$ where $m_\mathrm{x}$ is the subgroup membership of x, the preference function of x in $\mathcal{T}_i$ is $f_{im_\mathrm{x}}(t), t \in \mathcal{T}_i$. In most cases, $f_{im_\mathrm{x}}(t)$ is a function that increases first and then decreases, which has a maximum point corresponding to the optimal outcome. When it comes to the real world, for a group of people who have the same interest, their watching time of a certain type of short-videos increases as the the exposure proportion of this category gets bigger before the optimal location being reached, which is then reduced when the optimal proportion has been exceeded. It is assumed that $f_{im_\mathrm{x}}(t)$ appears like a mountain-shape curve and has a unique maximum point. However, $f_{im_\mathrm{x}}(t)$ can be multi-modal and contain several local optimums in practice due to the influence of noises. Therefore, the optimal treatment obtained by taking the maximum of the preference function can be extremely biased when there exists some outliers. For instance, a user can have an unusually long watching time if he falls asleep with the APP open although the exposure proportions randomly assigned to him are far from the mean optimal values of the subgroup he belongs to. Therefore, we need to design a robust strategy to make the estimated optimal point as close as possible to the population level optimal one.

To search for the optimal treatment, we introduce the cumulative preference function for the $i$-th treatment and subgroup $\mathcal{G}_k$ as

$$F_{ik}(t_0) = \int_{t\le t_0} f_{ik}(t)dt, t \in \mathcal{T}_i \tag{2}$$

which transfers $f_{ik}(.)$ into $F_{ik}(.)$ and is similar to integrating the probability distribution function (p.d.f.) to get the cumulative distribution function (c.d.f.). Instead of looking for the maximum of the preference function which is sensitive to outliers, we resort to the cumulative preference function at a higher level. Since the treatment closer to the optimal one are more likely to have a higher outcome, the area enclosed by a neighborhood near the optimal treatment is referring to the the location on the c.d.f. curve with the steepest slope, as Figure 3 shows. Hence we can reasonably infer that the mean of the $F_{ik}(.)$'s on the right hand side of the optimal point is significantly larger than that on the left hand side. Following this idea, we can estimate the optimum of the $i$-th treatment for subgroup $\mathcal{G}_k$ as

$$\begin{aligned}\hat{t^*_{ik}} &= \underset{t_i\in\mathcal{T}_i}{argmax} D(\mathcal{G}_k, t_i) \\ &\triangleq \underset{t_i\in\mathcal{T}_i}{argmax} \frac{1}{max_i - t_i}\int_{t>t_i} F_{ik}(t)dt - \frac{1}{t_i - min_i}\int_{t\le t_i} F_{ik}(t)dt,\end{aligned} \tag{3}$$

where $min_i$ and $max_i$ are the minimum and maximum of $\mathcal{T}_i$, respectively. In practice, the true $f_{ik}(.)$ and $F_{ik}(.)$ are unknown, and can be empirically approximated through discretization using observed data, and thus we need to find $t_i$ that maximizes the difference of empirical means of $F_{ik}(t)$ for $t > t_i$ and for $t \le t_i$ according to (3). It is noted that the maximum of $f_{ik}(.)$ is corresponding to the $F_{ik}(.)$ with the largest derivative (or slope). Our strategy is robust to some potential local violent fluctuations which result in extremely large first derivatives, and thus has the ability to control the estimation bias caused by outliers.

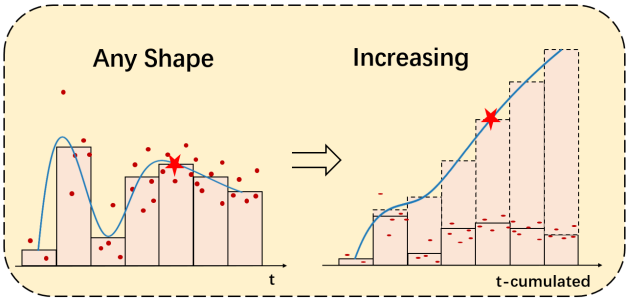



**Figure 3: The original preference function $f_{ik}(.)$ (left) and the transformed cumulative preference function $F_{ik}(.)$ (right). The optimal point is marked by a star.**

### 3.3 Splitting criterion for multi-dimensional continuous treatment

With all the prerequisites established in Sections 3.1 and 3.2, we can now introduce the splitting criterion for the more general multi-dimensional continuous treatment considered in this paper, i.e. $\mathrm{t} \in \mathcal{T}$ as follows,

$$\begin{aligned} &\max_{\mathcal{R}_l,\mathcal{R}_r} \sum_{i=1}^{K} (\mathcal{P}_i(\mathcal{R}_l) + \mathcal{P}_i(\mathcal{R}_r)) \\ &\triangleq \max_{\mathcal{R}_l,\mathcal{R}_r} \sum_{i=1}^{K} (|\mathcal{R}_l| D^2(\mathcal{R}_l, t^{*\hat{}}_{i\mathcal{R}_l}) + |\mathcal{R}_r| D^2(\mathcal{R}_r, t^{*\hat{}}_{i\mathcal{R}_r})), \end{aligned} \tag{4}$$

where $\mathcal{R}_l$ and $\mathcal{R}_r$ represent the samples in the left and right child nodes at a certain split. (4) reweights child nodes based on the total numbers of samples, and makes those with more data contribute more to the splittings. The estimated optimal treatment for each child node is included to encourage the splitting process to recover the underlying partition of the feature space. Moreover, an estimation of the optimal treatment can be obtained at each splitting instead of only in the final leaf nodes.

When the constraint $\sum_{i=1}^{K} t_i = 1$ is imposed where $t_i$ is the exposure proportion of the $i$-th video category, we modify the splitting criterion in (4) to make $\sum_{i=1}^{K} t^{*\hat{}}_{i\mathcal{R}_l} = 1$ and $\sum_{i=1}^{K} t^{*\hat{}}_{i\mathcal{R}_r} = 1$ hold for the child nodes $\mathcal{R}_l$ and $\mathcal{R}_r$, respectively. For the child node $l$, at each split, we take a random permutation of $\{1, \cdots, K\}$, denoted as $\{p_1, \cdots, p_K\}$, and find a threshold $k'$ such that $\sum_{i=1}^{k'-1} t^{*\hat{}}_{p_i\mathcal{R}_l} \leq 1$ and $\sum_{i=1}^{k'} t^{*\hat{}}_{p_i\mathcal{R}_l} > 1$. Besides that, we also let $t^{*\hat{}}_{p_{k'}\mathcal{R}_l} = 1 - \sum_{i=1}^{k'-1} t^{*\hat{}}_{p_i\mathcal{R}_l}$, and $t^{*\hat{}}_{p_i\mathcal{R}_l} = 0$ for $i > k'$. Similar procedures are performed to the other child node $r$ using the threshold $k''$. Thus, the modified splitting criterion is given by:

$$\max_{\mathcal{R}_l,\mathcal{R}_r} \sum_{i=1}^{k'} \mathcal{P}_{p_i}(\mathcal{R}_l) + \sum_{i=1}^{k''} \mathcal{P}_{p_i}(\mathcal{R}_r). \tag{5}$$

The formulation in (5) only involves training data thus may lack sufficient generalization power to perform well on unseen data. To increase the generalization ability of our method, we further split the original training set to get a validation set and control the discrepancy between them. We let $\lambda$ be a hyperparameter to regularize the validation loss, and derive the final form of the splitting criterion we use in this work,

$$\max_{\mathcal{R}_l,\mathcal{R}_r} \sum_{i=1}^{k'} (\mathcal{P}_{p_i}(\mathcal{R}_l) - \lambda L_{p_i}(\mathcal{R}_{l,val})) + \sum_{i=1}^{k''} (\mathcal{P}_{p_i}(\mathcal{R}_r) - \lambda L_{p_i}(\mathcal{R}_{r,val})). \tag{6}$$

where $L_{p_i}(\mathcal{R}_{j,val}) = |\mathcal{R}_{j,val}||(D(\mathcal{R}_{j,val}, t^{*\hat{}}_{i\mathcal{R}_{j,val}}) - D(\mathcal{R}_{j,tr}, t^{*\hat{}}_{i\mathcal{R}_{j,tr}}))|$. $\mathcal{R}_{j,tr}$ and $\mathcal{R}_{j,val}$ here represent the training set and validation set for the child node $j$, where $j \in \{l, r\}$. Note that the estimated optimal values of some of the treatments filtered by the threshold $k'$ or $k''$ are automatically set to 0 using the above searching strategy which is not reasonable in practice. To address this issue, we propose a ensemble method to make all categories have a non-zero possibility to be exposed, whose details will be discussed in the following section.

### 3.4 Forest Ensemble

Although a single tree introduced in Section 3.3 can be used to search for the optimal policy, it lacks sufficient robustness when being applied in practice. The permutations of the different video categories used to generate the treatment vector may lead to extreme unfairness considering the design of (5). The categories among the top of the orderings are more likely to be assigned a positive exposure proportion and one treatment dimension will be zero if those before it have already been summing to 1. Therefore, we propose a Forest Ensemble approach to ensure that all the video categories are equally treated, where different permutations of the video categories are used each time. Similar to other forest models, the ensemble approach can improve both the diversity and the robustness of the proposed method, while a larger forest scale is required to balance the feature selections and the treatment ordering.

### 3.5 Computation Acceleration

Due to the complex design of MDP$^2$ Forest, including the 4 nested loops and the forest ensemble, its empirical efficiency can not be guaranteed without any further optimization. To largely reduce the computation time of MDP$^2$ Forest , we introduce two training strategies: Quantile Sketch and Quantile Interval Average.

**Quantile Sketch.** Employing a naive numerical optimization procedure to find the optimal values for both the features and the treatment requires a lot of computation time considering the large searching space to be explored. Our proposed Quantile Sketch approach sorts the sample-level $x_{ij}$'s and $t_{ij}$'s, and then calculates the quantile values for each of them. With the selected quantiles, we can effectively approximate the distribution of the raw data. In practice, using 20 quantiles instead of the original 50,000 samples increase the running speed by over one million times.

**Quantile Interval Average.** Quantile Sketch can not only help reduce the training time, but also benefit the estimation of the cumulative preference function provided by 3.2. Suppose we have $p$ quantiles $\{Q_1, Q_2, \cdots, Q_p\}$ for the $j$-th dimension of the treatment vector in subgroup $\mathcal{G}_k$, the preference function $f_{jk}(t_i)$ of the treatment $t_{ij}$ between $Q_l$ and $Q_{l+1}$ can be approximated by the mean value of the interval. Under this setting, the cumulative preference function valued at $t_{ij}$ is the weighted sum of the interval means

before $t_{ij}$, i.e.

$$F_{jk}(t_0) = \frac{\sum_{i=1}^{l} \bar{y}_{Q_i} \cdot (Q_{i+1} - Q_i)}{(Q_{l+1} - Q_1)}, \quad Q_l < t_0 \leq Q_{l+1} \tag{7}$$

where $\bar{y}_{Q_i}$ is the mean value of the intervals determined by the quantile locations $Q_i$ and $Q_{i+1}$. Therefore, we only need to compute $F_{jk}(Q_1), F_{jk}(Q_2), \cdots, F_{jk}(Q_p)$ in one round of treatment value iteration, which greatly reduces the computational complexity of the cumulative preference function.

# 4 EXPERIMENTS

In this section, we empirically compare the proposed MDP$^2$ Forest with other causal inference or policy learning methods in two datasets. One is simulated using randomly generated features and treatments, which mimics a randomized trial with little co-variance conducting to different populations. The other is generated using the real-world business data and has co-variance between features and treatments.

One big challenge for the synthetic experiment is that we don't know the real optimal treatments of the users. To avoid very high trial-and-error costs by doing online verification, we use policy-based generated outcomes instead of the real-world ones to evaluate all the compared methods. Based on the fact that better strategies lead to higher outcomes, we take the mean regret as the evaluation metric for both datasets according to the treatments people receive.

For each experiment, the whole dataset is randomly divided into a training set (40,000 samples) and a test set (10,000 samples). For those methods that need a validation set, the training set is further split to get a validation set with 10,000 samples.

MDP$^2$ Forest has two different kinds of splitting criteria described in (5) and (6), which are respectively named by MDP$^2$F and MDP$^2$F-PV (Penalty on Validation loss) as a distinction. We compare the two variants of MDP$^2$ Forest with five different existing methods that are applicable to continuous-treatment problems, including Double Machine Learning with the power of treatments (DML), Dose Response Network (DRNet), Varying Coefficient Network (VCNet), continuous Off-Policy Evaluation (OPE)[12], and Optimization over Continuous and Multi-dimensional Decisions (OCMD)[3].

All the compared methods are originally designed for causal inference or policy learning problems. For OPE, a more general version of the off-policy evaluator using a multi-dimensional Epanechnikov kernel is employed in this work although the original paper of OPE [12] only pays attention to the one-dimension case. For OCMD, we use a linear model and a Lasso model as the first-stage predictive models for the simulated and synthetic experiments, respectively. Then, the second-stage optimization is conducted by solving a second-order conic optimization problem. For the uplift models, i.e. DML, DRNet and VCNet, we take the peak of the treatment-response curve as the optimal exposure proportion for each category.

In particular, it is worth noting that the uplift methods can't handle the multi-dimensional case with the constraint in the total quantity. Therefore, some adjustment is required to make them work. A simple way used by the industry is to calculate the treatment effect for one category each time and then do some normalization to make the sum-to-one constraint satisfied. For example, the optimal treatment vector obtained by these methods in a four-dimensional case is $(0.8, 0.2, 0.6, 0.4)^T$ without doing any adjustment. And the final exposure proportions will be $(0.4, 0.1, 0.3, 0.2)^T$ after normalization.

## 4.1 Evaluation Metrics

We introduce two metrics, Mean Regret (MR) and Mean Treatment Square Error (MTSE), to evaluate the compared methods.

*4.1.1 MR (Mean Regret).* Regret is a classical metrics for policy evaluation, which is the mean difference between the predicted optimal outcome and the ground truth, i.e.

$$MR = \frac{1}{n} \sum_{i=1}^{n} (y(\pi^*(x_i)) - y(\hat{\pi^*}(x_i)) \tag{8}$$

where $n$ is the total number of observations, and $y(\pi^*(x_i))$ is the theoretically optimal outcome of the $i$-th sample. In practice, MR indicates how much revenue each method can bring, and a smaller MR represents a better recommendation strategy.

*4.1.2 MTSE (Mean Treatment Square Error).* To effectively show how close the predicted optimal treatment provided by each compared method and the true one are under the multi-dimensional case, we introduce the MTSE as follows,

$$MTSE = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} (\pi_j^*(x_i) - \hat{\pi_j^*}(x_i))^2$$

where $n$ is the sample size, $m$ is the total number of treatment dimensions, and $\pi_j^*(x_i)$ is the predicted optimal exposure proportion for the $j$-th category and the $i$-th user. A smaller MTSE represents a better prediction result.

## 4.2 Simulation Study

We first generate simulation data to conduct an intuitive comparison between different methods, which is a simplification of the real-world problem. The necessities of the simulation study are in two aspects. First, it simulates the randomization experiment by including counterfactual examples. Some of the compared methods are expected to perform well in this case, which allows us to test whether these methods can identify the heterogeneity within the data and demonstrate the advantage of our method. Second, the simulated observations in each group are assumed to have the same optimal treatments and outcomes, which helps us to quantify the closeness between the learned optimal policy and the ground truth using the group level mean values.

We simulate six features for each training sample to represent users' characteristics, including sex, age, education background, city level and two behavior characteristics $x_\alpha$ and $x_\beta$. All the six features are randomly and independently generated. The treatment has a length of six, which represent the exposure proportions of the six different video categories, including films, food, games, life, outdoors and beauty. We simulate the number (or the duration if necessary) of videos in each category exposed to each user, and then calculate the video exposure proportions as the observed treatment.

Based on users' characteristics, we divide the whole population into 8 groups, and assign the same group of users the same optimal video exposures. The best video exposure proportions for each group and the underlying generating mechanisms are shown in Table 2 of the Supplement. As for the outcome, the active duration of users is simulated under the assumption that the outcome will be higher if the assigned treatments are closer to the best ones.

| Approach | Film | Food | Game | Life | Out. | Bea. |
|---|---|---|---|---|---|---|
| True Optimal | .25 | .25 | .35 | .05 | .05 | .05 |
| MDP$^2$F | .246 | .223 | .403 | .048 | .040 | .040 |
| MDP$^2$F-PV | .208 | .208 | .344 | .063 | .054 | .122 |
| DML | .175 | .327 | .455 | .016 | .003 | .025 |
| DRNet | .158 | .139 | .279 | .145 | .139 | .139 |
| VCNet | .252 | .059 | .137 | .252 | .173 | .128 |
| OPE | .173 | .193 | .292 | .109 | .140 | .092 |
| OCMD | .202 | .129 | .366 | .147 | .091 | .065 |

**Table 1: The mean optimal exposure proportion of Group 1 predicted by each method. MDP$^2$F significantly outperforms the others.**

As Table 1 shows, MDP$^2$F and MDP$^2$F-PV significantly outperform the other five with the predicted optimal treatment closer to the true optimal one in each category. Although the overall distribution estimated by DML is similar to the true one, the quantitative gaps between the predictions and the ground truths are significantly larger in all the six categories than those obtained by our method. DRNet and VCNet cannot learn the pattern of the exposure proportions, i.e. there is no obvious difference between the first three treatments and the latter ones. Moreover, the predicted optimal treatments vary a little across different categories, which indicates that the two methods cannot identify the heterogeneity.

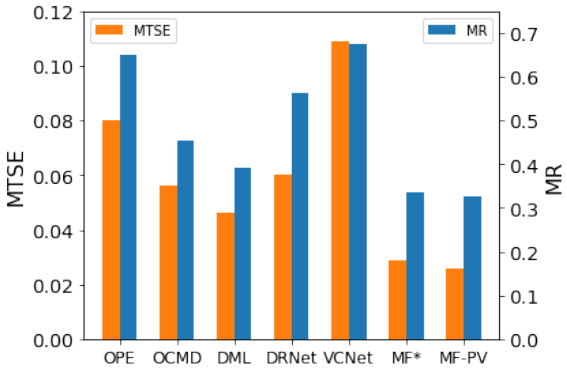



**Figure 4: MTSE and MR on the simulated data by the seven compared methods. * MF is short for MDP$^2$ Forest**

The evaluation results in terms of MR and MTSE are presented in Figure 4. Both MDP$^2$F and MDP$^2$F-PV obtain much lower MR and MTSE than any other method. MDP$^2$F-PV performs a little bit better than MDP$^2$F since it corrects the training biases caused by the split using a validation set. This result proves that adding a penalty term is meaningful, although the number of training samples is reduced and some extra time cost is incurred. DML, DRNets and VCNets perform much worse than our method due to the dimension normalization and their weakness in distinguishing the within-data heterogeneity. The use of local kernels in OPE does not fully address the curse of dimensionality, which results in the poor performance of OPE in the multi-dimensional setting.

## 4.3 Synthetic Study

Considering the low complexity and randomization design of the simulation study, we do a synthetic experiment in this part to further explore the empirical advantage of our method when applied in the real-world industry. All the compared methods are applied on a synthetic dataset to determine the optimal category proportions in short video recommendations.

No real-world datasets can be directly used to do policy evaluation since counterfactual outcomes are not fully observed in practice. Therefore, we employ a data generation procedure to recover the counterfactuals using observed data. Specifically, $\mathrm{x}_i$ represents the 20-dimensional user characteristics (e.g. age, gender, active institute and active days). The treatment $\mathrm{t}_i$ is a 10-dimensional vector to represent the exposure proportions of 10 categories of short-videos. The outcome $y_i$ is the watching time of the $i$-th user. The details of the data generation procedure is described in B.2 of the Supplement. For evaluation, the optimal watching time of a user in subgroup $c$ is obtained by plugging the optimal treatment (peak location) $\mathrm{t}^*$ into the counterfactual generating function $\bar{y}_c p_c(\mathrm{t})$. Figure 5 presents the MR and MTSE results on the test set by each of the seven methods. In this case, MDP$^2$F still achieves the best performance and its advantage becomes even larger than in the simulation study. Moreover, MDP$^2$F-PV also outperforms MDP$^2$F in both MR and MTSE, which validates the effectiveness of penalizing the error on the validation set when training the tree model. The kernel-based method OPE performs much worse than in the simulation study due the higher dimensionality of the treatment vector. The optimal bandwidth selection strategy suggested by [12] is not suitable to the synthetic dataset, thus we choose the optimal bandwidth for OPE through grid search evaluated on the validation set.

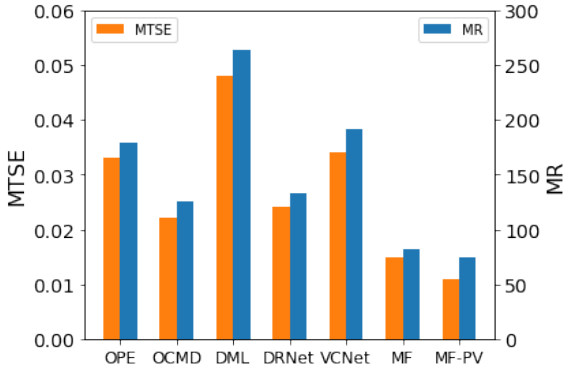



**Figure 5: MTSE and MR on the simulated data by the seven compared methods.**

For OCMD, the linear model is applicable for the first-stage predictive task and would result in a bad performance, which is replaced by a Lasso model which fits more for the high dimension scenario. DML performs the worst in the synthetic study due to the underlying complex correlation within the dataset and the high dimension of the treatment space.

The synthetic dataset can be regarded as an easy-evaluating equivalent of the real-world industrial problem. A small MTSE indicates that the video recommended to users can better meet their hobbies and diversified needs, and the reduction of MR means that the user's online time is closer to the limit that the recommendation strategy can bring. Our method shows a significant superiority than any other in the field of policy learning, and thus promise great potential to the industry.

### 4.4 Hyperparameter Analysis

We do some further analysis to show how different number of trees bagged by MDP$^2$F-PV and MDP$^2$F affect the model performance on the synthetic dataset in terms of MR and MTSE. Figure 6 (Appendix C) visualizes the changes of MR and MTSE, against the number of trees in the forest, from 1 to 300. We observe that the two evaluation metrics keep decreasing as the total number of trees increases, and eventually reach a convergence when over 250 trees are included. Moreover, applying a penalty on the validation loss helps converge faster and perform much better than the original model given limited trees.

It's worth noting that the two evaluation metrics MR and MTSE are not available in practice since the counterfactual outcomes are unobserved and thus the true optimal treatment can not be obtained. An alternative approach is to decide when the estimated optimal treatments on validation set substantially converges as the number of trees increases.

## 5 RELATED WORK

The main target of short-video recommendation problems is to develop optimal personalized treatments or optimal contextual policy, analyzing their significance and deficiency. Most existing literature in this field focuses on discrete treatment space, where limited possible treatments are considered [2, 4]. However, extending them to the continuous case is not trivial. Directly solving with the continuous policy optimization using naive discretization strategies may result in poor empirical performance [30]. A popular and intuitive two-step regression-based method Q-learning [18, 20] models the outcome as a function of treatments and subject features, seeking for the optimal treatment that maximizes the function. The main disadvantage of Q-learning is due to the issue of model misspecification, especially under the multi-dimensional continuous scenario where the performance is highly sensitive to the accuracy of the regression model.

Although many policy learning methods for discrete treatments are available, the methods for continuous policy learning received relatively less attention with a few exceptions. Demirer et al. [9] studies continuous policy evaluation through a semi-parametric approach where part of the value function satisfies a known parametric form. However, such prior knowledge might not be available in practice because of the complex mechanisms of the real-world problems. Chen et al. extends outcome weighted learning method to personalized continuous dose findings under a randomized trial design instead of on the observational data [5]. Kallus and Zhou propose a non-parametric kernel-based method that can be extended to multi-dimensional treatments using multi-dimensional kernel functions [12], which suffer from the curse of dimensionality and perform poorly when the dimension of treatment is high. Zenati et al. addressed the continuous policy learning problem using counterfactual risk minimization (CRM) based on a joint kernel embedding of features and treatments [30]. The difference between our work and [30] lies in that we do not assume known propensities of the logging policy. Bertsimas and McCord proposed an uncertainty penalization approach that can handle continuous and multi-dimensional treatments [3]. However, [3] requires a predictive model to estimate the cost (outcome) that may have similar problems as in Q-learning, and penalization-based methods usually bring a large computational burden. Furthermore, these policy learning methods are based on numerical optimization, which is sensitive to tuning parameters, and thus, usually unstable for online deployment.

Policy optimization is closely related to estimation and inference of heterogeneous treatment effects (HTE), which is of great interest in the area of causal inference. Most of the studies in this area focus on making inferences on the causal effect of a treatment based on observational data [24]. An important method for HTE inference is double/debiased machine learning (DML) proposed by [8]. DML allows arbitrary predictive machine learning algorithms for two sub-tasks and maintains favorable properties of the estimator, thus reducing the risk of model misspecification compared with Q-learning and other parametric methods. Schwab et al. proposes a neural network approach called DRNets to estimate individual dose-response curves for any number of treatments [21]. Nie et al. improves the continuity by introducing the varying coefficient model to the network, and put forward VCNet[19]. In fact, the problem of HTE involves too much to be estimated as the causal effect for every possible treatment. Policy optimization can be performed after obtaining counterfactual estimates of causal effects. However, as noted by Tanimoto et al. [22] and our empirical experiments, there is a gap between the decision-making performance and the estimation accuracy of causal effects. The decision performance is not guaranteed even when the estimation accuracy has already reached a high level.

Trigger, proposed by Tran and Zheleva [23], is an extension of HTE. Trigger is a personalized threshold of the treatment needed to be reached in order to obtain a significant effect. For continuous treatments, the trigger can be found by iterating over all possible treatment values and serve as the threshold to maximize the causal effect when splitting the data into treatment and control groups. However, this method cannot be applied to multi-dimensional continuous case.

## 6 CONCLUSIONS AND DISCUSSIONS

In this paper, we propose a novel short-video recommendation method called MDP$^2$ Forest, which models the exposure proportions of different video categories as a multi-dimensional continuous treatment with a sum-to-one constraint. We improve the partition measures by discretizing the continuous treatment values and do the policy searching using the cumulative preference function. And the splitting criterion is an aggregation function of different treatment dimensions. We also explain how these designs fit the specialty of the short-video recommendation problem. The experimental results show that our method is much better than some state-of-the-art methods in this specific scenario.

There also exist some important adjustments and future directions. First, according to various business scenarios, the constraint should be modified, and the stacking method of optimal treatments needs to be adjusted. Second, instead of the "greedy strategy", we can apply some advanced methods to find the global optimum. Third, the splitting criterion may be improved to reduce the bias.

## 7 ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (12001356), Shanghai Sailing Program (20YF1412300), "Chenguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission, Open Research Projects of Zhejiang Lab (NO.2022RC0AB06), Innovative Research Team of Shanghai University of Finance and Economics (2020110930), Fundamental Research Funds for the Central Universities.

We thank Tencent PCG for providing computing resources and research data. Also thanks to Shikai luo, Ge Song, Harry Hu and Yifan Sun for their helpful suggestions.

## REFERENCES

[1] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
[2] Susan Athey, Stefan Wager, et al. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 78, 2017.
[3] Dimitris Bertsimas and Christopher McCord. Optimization over continuous and multi-dimensional decisions with observational data. *arXiv preprint arXiv:1807.04183*, 2018.
[4] Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru. Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.
[5] Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521, 2016.
[6] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ArXiv*, abs/2010.03240, 2020.
[7] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. Autodebias: Learning to debias for recommendation. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.
[8] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.
[9] Mert Demirer, Vasilis Syrgkanis, Greg Lewis, and Victor Chernozhukov. Semiparametric efficient policy learning with continuous actions. *arXiv preprint arXiv:1905.10116*, 2019.
[10] Weihao Gao, Xiangjun Fan, Chong Wang, Jiankai Sun, Kai Jia, Wen Xiao, Ruofan Ding, Xingyan Bin, Hui Yang, and Xiaobing Liu. Learning an end-to-end structure for retrieval in large-scale recommendations. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.
[11] Hao Jiang, Wenjie Wang, Yin wei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.
[12] Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251. PMLR, 2018.
[13] Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. Semi: A sequential multi-modal information transfer network for e-commerce micro-video recommendations. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.
[14] Yongqi Li, Meng Liu, Jianhua Yin, C. Cui, Xin-Shun Xu, and Liqiang Nie. Routing micro-videos via a temporal graph-guided recommendation system. *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.
[15] Haochen Liu and Xiangyu Zhao. Self-supervised learning for alleviating selection bias in recommendation systems. 2021.
[16] Shan Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. User-video co-attention network for personalized micro-video recommendation. *The World Wide Web Conference*, 2019.
[17] Yiyu Liu, Qian Liu, Yu Tian, Changping Wang, Yanan Niu, Yang Song, and Chenliang Li. Concept-aware denoising graph neural network for micro-video recommendation. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.
[18] Erica EM Moodie, Nema Dean, and Yue Ru Sun. Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6(2):223–243, 2014.
[19] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.
[20] Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.
[21] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020.
[22] Akira Tanimoto, Tomoya Sakai, Takashi Takenouchi, and Hisashi Kashima. Regret minimization for causal inference on large treatment space. In *International Conference on Artificial Intelligence and Statistics*, pages 946–954. PMLR, 2021.
[23] Christopher Tran and Elena Zheleva. Learning triggers for heterogeneous treatment effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5183–5190, 2019.
[24] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
[25] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *ICML*, 2019.
[26] Yin wei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.
[27] Dong Yao, Shengyu Zhang, Zhou Zhao, Wenyan Fan, Jieming Zhu, Xiuqiang He, and Fei Wu. Modeling high-order interactions across multi-interests for micro-video reommendation. In *AAAI*, 2021.
[28] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.
[29] Jiangxing Yu, Hong Zhu, Chih-Yao Chang, Xinhua Feng, Bowen Yuan, Xiuqiang He, and Zhenhua Dong. Influence function for unbiased recommendation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.
[30] Houssam Zenati, Alberto Bietti, Matthieu Martin, Eustache Diemert, and Julien Mairal. Counterfactual learning of stochastic policies with continuous actions: from models to offline evaluation. 2021.
[31] Qi Zhang, Longbing Cao, Chongyang Shi, and Liang Hu. Tripartite collaborative filtering with observability and selection for debiasing rating estimation on missing-not-at-random data. In *AAAI*, 2021.
[32] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

# A PSEUDO-CODE FOR PARTITION MEASURE

**Algorithm 1** Partition Measure

**function** PARTITIONMEASURE($\mathcal{D}_\ell$)
  $\{\theta_\ell\} \leftarrow \{\emptyset\}$; {denote the multi-dimensional MDPP}
  $\mathcal{B}_\ell \leftarrow 0$; {the best partition measure}
  $\mathbb{R} \leftarrow 1${the amount of available resources}
  $\mathcal{T} = \{t_{i1}, \cdots, t_{iK} : t_{ij} \in \mathcal{D}_\ell\}$
  **for** each $j \in \{1, \cdots, K\}$ **do**
    $O_j \leftarrow -\infty$
    **for** each quantile $t_{ij}$ of $\mathcal{T}_j$ {where $i \in \{1, \cdots, 20\}$} **do**
      $\mathcal{D}_{\ell_2} = \{(\mathbf{X}_{ij}, Y_{ij}, T_i) : T_{ij} \geq t_{ij}\}$,
      $\mathcal{D}_{\ell_1} = \{(\mathbf{X}_{ij}, Y_{ij}, T_i) : T_{ij} < t_{ij}\}$
      temp = $F(\mathcal{D}_{\ell_2} \cup \mathcal{D}_{\ell_1})$
      **if** temp > $O_j$ **then**
        $O_j, \theta_j \leftarrow temp, t_{ij}$
      **end if**
    **end for**
    **if** $\mathbb{R} < \theta_j$ **then**
      $\theta_j \leftarrow \mathbb{R}$
    **end if**
    $\mathbb{R}, \mathcal{B}_\ell \leftarrow \mathbb{R} - \theta_j, \mathcal{B}_\ell + O_j$
    **if** $\mathbb{R} == 0$ **then**
      **break**
    **end if**
  **end for**
  **return** $\mathcal{B}_\ell, \{\theta_\ell\}$

# B DESIGN OF THE DATASETS

## B.1 Simulated Data

The assumptive best video exposure proportions of each group are shown in Table 2.

| Feature | | | Best Exposure Proportion | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age | Edu | $x_\alpha$ | Film | Food | Game | Life | Out. | Bea. |
| $< 45$ | $\geq 2$ | $\geq 0.5$ | .25 | .25 | .35 | .05 | .05 | .05 |
| $< 45$ | $\geq 2$ | $< 0.5$ | .25 | .05 | .35 | .05 | .05 | .25 |
| $< 45$ | $< 2$ | $\geq 0.5$ | .05 | .25 | .35 | .05 | .25 | .05 |
| $< 45$ | $< 2$ | $< 0.5$ | .05 | .05 | .35 | .05 | .25 | .25 |
| $\geq 45$ | $\geq 2$ | $\geq 0.5$ | .25 | .25 | .05 | .35 | .05 | .05 |
| $\geq 45$ | $\geq 2$ | $< 0.5$ | .25 | .05 | .05 | .35 | .05 | .25 |
| $\geq 45$ | $< 2$ | $\geq 0.5$ | .05 | .25 | .05 | .35 | .25 | .05 |
| $\geq 45$ | $< 2$ | $< 0.5$ | .05 | .05 | .05 | .35 | .25 | .25 |

**Table 2: The optimal video exposure proportion of each group. The population is divided into 8 groups according to the features. The optimal exposure proportion is heterogeneous across different groups and homogeneous within a group.**

Specifically, the formula of constructing the response is

$$y = \sum_{i \in \{age, edu, x_\alpha\}} (-|t_{i1} - B_{i1}| - |t_{i2} - B_{i2}| + B_{i1}^2 + B_{i2}^2) + \epsilon$$

where $B_{ij}$ is the best exposure proportion shown in Table 2 and $t_{ij}$ is the corresponding treatment of the sample.

In our data generation process, there are 3 features that affect the response and each feature can decide the value of $B_{i1}$ and $B_{i2}$. For example, if one sample has an educational background that is not less than 2, 0.25 and 0.05 will be the best exposure proportion of film and outdoors respectively, which can also be described as $B_{edu1} = 0.25$ and $B_{edu2} = 0.05$ in the above formula. $\epsilon$ is an additional item that makes sure the responses of all the responses are positive. Under the setting of the absolute value, the T-Y curve of a single treatment should have a straight mountain shape.

## B.2 Synthetic Data

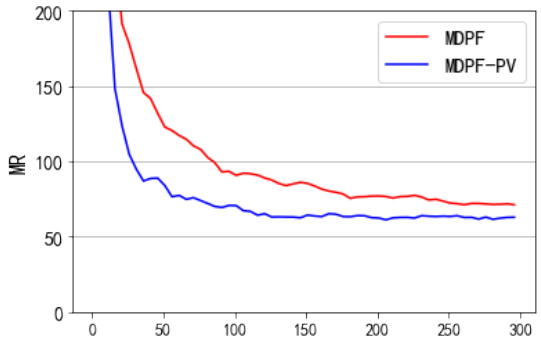
As one of the popular short video apps, Tencent Wesee has a DAU (the number of daily active users) of more than 30 million, which enables us to collect a sufficient amount of data. Consider a dataset collected from an unknown logged policy $\mathcal{D} = \{(\mathrm{x}_i, \mathrm{t}_i, y_i(\mathrm{t}_i))\}_{i=1}^n$ collected by Wesee. The reason why the policy is unknown is that the proportion of each category of video given by the historical policy for each user is implicitly determined by the recommendation system, which contains multiple complex procedures. The dataset is anonymized so that private information is not disclosed. The context $\mathrm{x}_i$ represents the 20-dimensional user characteristics for each user (e.g. age, gender, active institute and active days). The treatment $\mathrm{t}_i$ represents the 10-dimensional vector of exposure proportions for 10 categories of videos. The outcome $y_i(\mathrm{t}_i)$ is the viewing time of the $i$-th user. The sketch of counterfactual generating process is shown as follows.

(1) Construct the undirected graph $G_x$ through performing KNN on $\{\mathrm{x}_i\}_{i=1}^n$, and then do Louvain clustering on $G_x$. The number of cluster is adaptively determined by Louvain clustering and unknown to us.
(2) For each cluster, construct a 10-dimensional multivariate normal p.d.f., denoted as $p_c(.)$, to model the causal relationship between the treatment and outcome for the $c$-th subgroup of users. The peak location of the normal curve is given by the $\frac{|\bar{\mathrm{t}_c} + \epsilon_c|}{\sum_{i=1}^{10} |\bar{t_{ci}} + \epsilon_{ci}|}$, where $\bar{\mathrm{t}_c}$ is the mean treatment vector of the $c$-th cluster, $\epsilon_{ci}$s are mutually independent and $\epsilon_{ci} \sim N(0, 0.05)$. Each element of the covariance matrix is given by the sample covariance of the corresponding pair of exposure proportions in this cluster.
(3) For each sample in the $c$-th cluster, the counterfactual outcome given a treatment $\mathrm{t}$ is $\bar{y}_c p_c(\mathrm{t})$, where $\bar{y}_c$ is the sample mean outcome of observed outcomes in the $c$-th cluster.

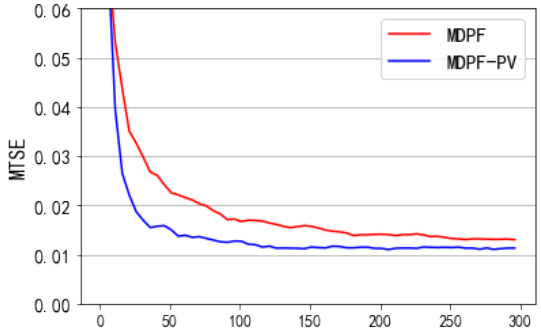
Given the counterfactual generating process, the synthetic dataset is $\mathcal{D}_{syn} = \{(\mathrm{x}_i, \mathrm{t}_i, \bar{y}_{c_i} p_{c_i}(\mathrm{t}_i))\}_{i=1}^n$, where $c_i$ is the subgroup membership of the $i$-th user. In our synthetic dataset, the features and treatments directly come from the observed real-world data, only the corresponding outcomes are generated from the above-designed process. The synthetic dataset basically preserves the distribution of the real-world data and any possible counterfactual can be obtained by the generating process. Therefore, the standard evaluation protocol as in the simulation study can be reasonably realized based on the synthetic dataset. To our knowledge, this is the first to formulate a counterfactual generating procedure with

multi-dimensional continuous treatment space, and can fit in different observed datasets with only minor changes. Additionally, the setup (peak location and covariance) of the normal curve can be modified to incorporate prior knowledge about the optimal treatment or some constraints, and here we add random noise to the mean treatment given by the historical recommendation system. Compared with the simulated dataset, the synthetic dataset has higher dimensional features and treatments, and a more complex relationship between $y$ and $(x, t)$, thus can be regarded as a higher-level task.

## C HYPERPARAMETER ANALYSIS




(a) MR




(b) MTSE

**Figure 6: Changes of MR and MTSE against the number of trees in the forest on the synthetic data by MDP$^2$F-PV and MDP$^2$F.**