Research Article Open Access

Jeroen Berrevoets*, Sam Verboven, and Wouter Verbeke

Treatment Effect Optimisation in Dynamic Environments

DOI: XXX

Abstract: Applying causal methods to fields such as healthcare, marketing, and economics receives increasing interest. In particular, optimising the individual-treatment-effect— often referred to as uplift modeling —has peaked in areas such as precision medicine and targeted advertising. While existing techniques have proven useful in many settings, they suffer vividly in a dynamic environment. To address this issue, we propose a novel optimisation target that is easily incorporated in bandit algorithms. Incorporating this target creates a causal model which we name an *uplifted contextual multi-armed bandit*. Experiments on real and simulated data show the proposed method to effectively improve upon the state-of-the-art. All our code is made available online at https://github.com/vub-dl/u-cmab.

Keywords: Bandit Algorithms, Uplift Modeling, Individual Treatment Effect

1 Introduction

Making individual-level causal predictions is an important problem in many fields. For example, *conditional average treatment effect* (CATE) predictions can be used to: prescribe medicine only when it causes the best outcome for a specific patient; advertise only to customers that were not going to buy otherwise; or grant subsidised job-training limited to those that find a job because of it[1].

While many methods exist to make CATE predictions, they fail to adapt through time. We believe this is a crucial problem in causal inference as many environments are dynamic in nature: patients could build a tolerance to their prescribed medicine; the initial marketing campaign could suffer from a competitor's product release; or the ever changing economic landscape could heavily influence new hires [4]. Even more so, existing models could introduce a bias in others once deployed, with detrimental performance as a result [5]. Deteriorating behaviour due to a dynamic environment is, in machine learning, also known as concept drift [6, 7].

We focus on the problem of optimising the treatment outcome based on estimates of the CATE, under a *strong ignorability assumption*. For example, if data on patients is collected by an expensive doctor, his wealthier patients—with potentially healthier diets —could greatly influence the outcome of their prescribed medicine. Strong ignorability then assumes that every confounder expressing influence on the outcome of a treatment is registered in the data, i.e., the patients diet [8–10]. This type of optimisation strategy is, in the literature, often referred to as CATE optimisation, or uplift modeling [11, 12].

Estimating the causal effect of a treatment contrasts classical learning as it is impossible to use the individual-level effect as a target in supervised learning. The reason for this is that we can only witness the outcome of exactly one treatment, leaving no other outcome— e.g., the outcome without treatment —to compare with. For example, it is impossible for a patient to both take medicine, and not take medicine simultaneously, for the doctor to compare

1 There exists some confusion in the treatment effects literature, where work concerning the *individual* treatment effect, is actually referring to CATE. For an explicit difference between these we refer to [1–3]

***Corresponding Author: Jeroen Berrevoets:** Department of Applied Mathematics and Theoretical Physics, University of Cambridge; E-mail: jeroen.berrevoets@maths.cam.ac.uk
Sam Verboven: Data Analytics Laboratory, Vrije Universiteit Brussel; E-mail: sam.verboven@vub.be
Wouter Verbeke: Faculty of Economics and Business, Leuven.AI, KU Leuven; E-mail: wouter.verbeke@kuleuven.be

the outcomes. This problem is known as the *fundamental problem of causal inference* [13]. The outcome of the alternative treatment is often termed a counterfactual outcome [14].

As a first naive attempt to create a dynamic causal inference model, one could try to employ an on-line learning method, e.g., an on-line random forest [15]. However, these methods require a target variable to adjust their current approximations, which— due to the fundamental problem of causal inference —is unobservable, rendering such on-line approaches useless. A second naive approach would be to use a change detection algorithm [7], initiating a retraining subroutine when necessary. As shown in our experiments, this approach performs poorly.

We take a fundamentally different approach than the naive strategies above as we reformulate CATE optimisation as a bandit problem [16]. Contrasting a standard supervised setting, a bandit setting does not assume a dataset a priori, allowing any required data to be collected when needed through active interaction with the data-generation process. Since there are bandits that learn continuously, they easily adapt to dynamic environments using a windowed estimation of their target [17], thereby offering solace against concept drift.

However, in order to redefine CATE optimisation as such, we must first define optimality in an CATE optimisation setting [11]. Furthermore, connecting CATE optimisation to literature on bandits — and potentially reinforcement learning — we postulate a more intelligent agent.

We have structured this paper as follows: first, we discuss recent studies related to ours (Section 2); next, we give a formal description of CATE and contextual multi-armed bandits (Section 3); then we give a detailed description of our proposed model (Section 4); and finally, we discuss the results of our experiments (Section 5).

2 Related work

Presented with a dataset, the vast majority of causal literature is concerned with inferring the causal effect one variable may have on another. Conceptually, there are two main perspectives one might take when estimating such an effect: (i) an effect may be described as a deterministic function[2], where the collection of functions between variables are structured in a graphical model [18]; or (ii) an effect may be described as the differential expectation of a variable's outcome given a alternative conditionals [8]. In this work, we adopt the latter. Of course, both approaches describe causal effect, but these separate views form a topic of controversy, where (ii) is stated to be subsumed by (i) [19, Section 4], or (i) is found to be sufficient to solve practical problems, particularly in economics [20].

While, a bandit perspective on (i) has already been explored in literature (see for example, [21–23]), providing such perspective in (ii) remains — to our knowledge — unexplored.

Potential outcomes. The main subject of analysis in the potential outcomes framework — introduced by [24] and [25] — is the unit-based response variable [19]. Typically, we reason about this response variable in a counterfactual manner: "all else the same, what would the outcome be had the treatment been different?". Naturally, this is a difficult question to answer, as in the data, we can only observe one of the potential outcomes, every other potential outcome is not observable [13].

Confounder bias. An issue in the potential outcomes framework, when working with a dataset, is that of confounder bias [10]. In essence, a naive estimate of the expected potential outcome will be biased equally to the treatment, leaving us two options [26]: (i) either the treatment is independent of $\mathbf{x}$, for example treatment is randomly assigned as in a randomised control trial; (ii) or we have to explicitly account for the bias in our dataset. As the first approach is costly and in some settings even unethical, a vast majority of research has focused on developing methods that adhere to the second approach in the static setting [10, 27–38], or in the longitudinal setting [39–46]. In this paper, we take a fundamentally different approach. Rather than working with a dataset, we train and collect data simultaneously.

CATE optimisation. Using an estimate of the CATE, one can optimise treatment selection in, for example, medicine [32, 47], marketing [48], or education [49]. While there is literature that identifies with optimisation specifically, the essence boils down to identifying a subpopulation with *high* treatment effect, and only treating those belonging to this subpopulation. Examples include include: uplift trees [48], multiple treatment decision trees

2 Usually paired with uncertainty baked into the parameters of this function.

[50], uplift random forests [51], and nearest-neighbours approaches [52]. Due to the lack of a ground truth label, evaluating these methods is non-trivial and is thus another line of active research [11, 53]. One particular evaluation metric is the Qini-curve [54], which we shall further discuss in Section 5.2.

Concept drift and change detection algorithms. While above literature provides good solutions to estimation, and eventually optimisation, dynamic environments can cause concept drift in a model. Concept drift is the phenomenon where the model approximated an out-of-date distribution with performance degradation as a result. [6] defines two types of concept drift: (i) gradual drift, where the underlying distributions of an environment tend to change gradually over time, e.g., when a patient builds up a tolerance for his medication; and (ii) sudden drift, when some event causes the underlying distributions to change suddenly, e.g., performance of a marketing campaign after the release of a competitor product. One way to counter concept drift with static methods, is to employ change detection algorithms. Change detection algorithms can set in motion a data collection and retraining routine when they notice performance degradation [7].

3 Background

Conditional average treatment effects. We are interested in an estimate of the net impact of a treatment $T \in \{0,1\}$ on a response $Y \in \mathbb{R}$ for an individual $\mathbf{x} \in \mathscr{X} \subseteq \mathbb{R}^d$. Specifically, we denote the net impact as the difference in conditional expectation,

$$\hat{u}(\mathbf{x}) \doteq \mathbb{E}[Y(T=1) - Y(T=0)|\mathbf{x}], \tag{1}$$

where $Y(T=1)$ is the potential outcome when the treatment is applied and $Y(T=0)$ is the potential outcome when it is not [8]. While we consider $Y(t)$ to be continuous, $Y(t)$ is often considered binary, denoting a responder or a non-responder [11]. If in practice we are interested in a binary $Y(t)$, we can simply consider the learned model a classification model [55]. In order to make Eq.(1) identifiable from data, we make assumptions deemed standard in CATE litarature.

Assumption 1 (Indentifiability assumptions.)**. Consistency** *When $Y = y$ and $T = t$, then $Y(T = t) = y$, meaning that an observed y in the data, corresponds to the potential outcome associated with t.* **Overlap** *for each individual, and each treatment we assume $0 < p(T|\mathbf{x}) < 1$.* **Conditional independence** *Given the individual, the potential outcomes are independent of the treatment, $Y(t) \perp\!\!\!\perp T|\mathbf{x}$.*

Above assumption is also known as the "strong-ignorability assumption", which warrants the potential outcomes to be identifiable. In principle, we assume here that every confounding variable is measured[3]. Further note that, *Overlap* and *Conditional independence* automatically holds when treatment is distributed randomly, as there is no dependence between the individual and treatment by definition [26, 57].

We use Eq.(1) to optimise treatment assignment. Given a high expected net impact ($\hat{u}$) we derive that an individual ($\mathbf{x}$) can be *caused* to respond positively to a treatment. CATE models are then used to identify a subpopulation of individuals with high $\hat{u}$[4]. By limiting treatment to this subpopulation we reduce over-treatment as we refrain from treating individuals indifferent to treatment ($\hat{u} = 0$) or worse, individuals that are averse to it ($\hat{u} < 0$)[5].

While still a problem in existing CATE optimisation techniques, strong ignorability is of less concern in our setting as bandits generally aim to operate given an unbiased estimate of their optimisation target. In fact, our experiments suggest this to be one of the major motivations to choose our causal agent over existing methods.

For an overview of various uplift methods and evaluation metrics we refer to [11] and [58].

3 With measured we mean that the variable is included in the domain of the dataset. The case of missing values of variables included in the domain for treatment effects is considered in [56]

4 The practice of identifying such a subpopulation is often referred to as *uplift modeling*.

5 Note that we assume high $Y(t)$ to be positive; if $Y(t)$ indicates, for example, tumour size, we simply flip the sign of $Y(t)$.

Contextual multi-armed-bandits. The bandit problem inherited its name from slot machines found in casinos [16]. Consisting of a set of arms, every arm of a K-armed bandit presents a response Y when pulled. For consistency, we shall refer to an arm as a treatment.

The objective of a CMAB algorithm is two-fold:

(i) Response prediction. A first objective is to predict what response a given treatment will yield. This is done through an expectation of a numerical representation of Y,

$$r(T=i,\mathbf{x}) \doteq \mathbb{E}[Y(T=i)|\mathbf{x}], \tag{2}$$

which draws immediate parallels to Eq.(1). Naturally, the treatment set can be much larger (than binary) without loss of generality, but is kept binary for consistency throughout this paper [59]. Further cementing the parallel between Eq.(1) and Eq.(2) is a major focus of this paper (another is aligning learning methods of CATE models and bandits).

As the response distributions are generally unknown (or assumed [60]), they are part of this learning objective. A direct consequence of this is the need for an estimation of r, which we shall denote $\hat{r}$. While methods exist to learn $\hat{r}$ from a dataset [61], it is generally the case to learn from one response at a time. As such, any regression method able to learn sequentially can be used to model $\hat{r}$ [17].

(ii) Optimal treatment selection. The second objective is to maximise Eq.(2) through treatment selection. Typically the CMAB algorithm identifies the optimal treatment by maximising $\hat{r}$ through T,

$$T^* = \arg\max_i \{\hat{r}(T=i,\mathbf{x})\}, \tag{3}$$

where T^* denotes the optimal treatment, given $\mathbf{x}$.

Randomly applying treatments ensures $\hat{r}$ to be unbiased, but contrasts Eq.(3). As such, the CMAB algorithm is required to properly balance: random treatment selection; and treatment selection through Eq.(3). This trade-off is often referred to as the exploration-exploitation trade-off [17]. Properly balancing exploration and exploitation is particularly important in environments that require online learning, e.g., dynamic environments subject to concept drift.

While many algorithms exist to balance exploration and exploitation, perhaps most intuitively is the ε-greedy algorithm. In ε-greedy, an optimal treatment T^* is chosen with probability $1-\varepsilon$ and a random T is chosen with probability ε. All responses are then used to formulate an estimate of $\hat{r}$. Another strategy is to sample treatment from a distribution which is then updated to be more optimal given additional data [60].

While ε-greedy offers satisfying results in a variety of environments, it is often not state-of-the-art. As our optimisation target is defined generally, any bandit algorithm can act causally as is illustrated with LinUCB [62] in our experiments. For a thorough overview of bandits we refer to [63] and [64]. In the following sections, we will argue that including appropriate costs in a bandit's objective, will yield the same objective used in uplift modeling approaches. Cost-constrained bandits are related to this (see for example, [65] and [66]), but not the focus of this work. Instead, our work suggests that, given the correct cost-definition, the aforementioned papers may be optimising a causal objective.

4 Method

The ultimate goal of our paper is to have a CMAB algorithm optimise the CATE in a dynamic environment. As bandits learn from every new data point they receive, they offer inspiration to build such an agent. Our strategy to achieve this goal is to define an appropriate optimisation target that lets any bandit make treatment decisions as if they were made on the basis of a causal CATE model.

We organise this section as follows: first we compare semantics between CATE and bandits, outlining some key differences between both methodologies; next we define an optimal treatment-threshold and illustrate its importance; finally we incorporate this threshold in an optimisation target for a bandit, creating the *uplifted contextual multi-armed bandit* (U-CMAB).

4.1 CATE versus contextual multi-armed bandits

We identify two main differences between bandits and ITE:

(i) Bandits and CATE models have different outputs. Given some context, CMAB algorithms output decisions directly. This is different in CATE optimisation where a model typically only outputs a ranking [11, 67]. Based on the estimated ITE[6], we compose a ranking of individuals where those with highest CATE are ranked above. Using this ranking, some external system (i.e., a practitioner) selects a threshold where all above receive treatment, and all below do not. With the U-CMAB, we want to remove this external decision system entirely and include CATE in the decision modeling literature.

(ii) Response optimisation is different from effect optimisation. The second difference concerns the objectives of bandits and CATE optimisation: Even when the optimal treatment T^* offers only a marginally higher response, it is chosen over all other T as bandits optimise for response only. This is in direct contrast with CATE where optimality has a different meaning. Only when a treatment offers *sufficient* effect — i.e., its estimated CATE is higher than an optimal threshold — it is considered optimal. The matter of how much effect is sufficient in order to apply treatment is dealt with in the following section. For the remainder of this paper we shall denote the optimal treatment in a bandit setting as T_b^* and an optimal treatment in CATE as T_u^*, where it is possible that $T_b^* \neq T_u^*$.

4.2 The optimal treatment-threshold

In this subsection, we discuss both the purpose and definition of the optimal treatment threshold.

The purpose of an optimal treatment threshold. We argue that the differences outlined above (in Section 4.1) are resolved through definition of the optimal treatment threshold. In CATE optimisation, a threshold is used to decide which individuals receive treatment, thus allowing CATE estimates to be translated in treatment decisions rather than argumentation, resolving the first difference. The second difference is accounted for by defining this threshold in function of effect.

Let $\tau \in \mathbb{R}$ be this optimal threshold with treatment selection in CATE optimisation as,

$$T_u^* = \mathbb{I}[u(Y, T, \mathbf{x}) > \tau], \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. When the support of $Y(t)$ is restricted to $[0,1]$, the support of τ is consequentially restricted to $[-1,1]$ as per Eq.(1).

We have illustrated Eq.(4) in Fig 1, with a simulated CATE in a two-dimensional context space [68]. With this simulation we visualise the effect of having an optimal threshold on CATE optimisation. A standard model would only estimate the ITE, i.e., coloured density plot. The bandit however, optimising for ITE, should use this estimated CATE to formulate the optimal region enclosed by τ, i.e., the region indicated by ITE $> \tau$. Whenever a context resides within this optimal region, it should receive treatment. As a causal agent is tasked with identifying these regions, knowing the correct threshold, τ, is paramount.

Definition of the optimal treatment threshold. Given an estimated ITE, $\hat{u}$ as in Eq.(1), a decision system typically defines τ heuristically through either: visual analysis [53]; or simple model evaluation metrics [54]. However, τ should consider the costs and potential benefits of a treatment as more costly treatments should cause higher effect before application. Therefore, we have defined τ as the difference between treatment costs as follows,

$$\tau = \psi_1 - \psi_0, \tag{5}$$

where the penalty — i.e., cost — of applying treatment $T = i$ is denoted $\psi_t : T \to \mathbb{R}^+$. Note that, if there are no costs associated with τ, we simply assign the treatment with highest estimated potential outcome (an example of this is a marketing e-mail campaign). Integrating τ into a bandit's objective, is then simply incorporating treatment costs into the unbiased treatment-arm estimates; discussed in more detail below. We find this an interesting result as it

6 Naturally, the ranking is based on CATE estimates. If one is solely interested in providing estimates, one could trivially not rank and select the individuals.

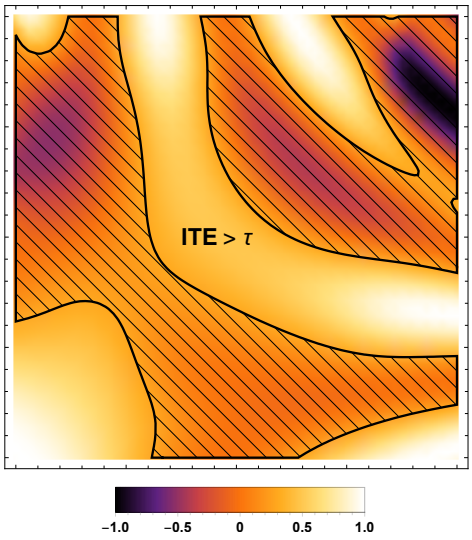


Figure 1. Difference between estimating, and optimising ITE. The region labelled ITE $> \tau$ is considered the optimal region given some τ; the shaded area indicates the suboptimal region. This figure illustrates the difference between mere CATE prediction— that is estimating the coloured density plot —and decision making based on a threshold, τ, and the estimated ITE— through the shaded area.

shows how closely related bandits are to CATE optimisation. The incorporation of costs has of course been explored previously in works such as [69].

From Eq.(4) we deduct that a lower τ will increase chances of applying treatment, i.e., a lower τ will enlarge the treatment area in Fig 1. As such, it makes sense that τ should increase when the penalty (ψ_t) associated with the treatment (t) is large. Defining τ as in Eq.(5) allows this penalty to increase, as long as a treatment's potential outcome increases accordingly. Naturally, we have $\psi_t < \mathbb{E}[Y(t)]$ in practice, as a treatment cannot be optimal when it costs more than its potential outcome [55].

Furthermore, with Eq.(5), we can extend Eq.(4) to a multiple treatment case by defining τ as a matrix T, with every component $\mathrm{T}_{ij} = \psi_i - \psi_j$ representing the required threshold for the CATE of $T = i$ against $T = j$ through Eq.(1). This implies that every treatment might require a different CATE before application due to a potentially different ψ_t. As such, assuming an optimal treatment exists, it corresponds with $\hat{u}(\mathbf{x}) > \mathrm{T}_{ij}$ for all $j \neq i$, including the control treatment $T = 0$. Note that this is a different strategy than most multiple treatment systems as they usually pick the treatment yielding most effect [4, 70].

4.3 The uplifted contextual multi-armed bandit

We stated that instead of relying on post-hoc heuristics to define τ, a definition such as in Eq.(5) helps us to define a bandit with an CATE optimisation target. In this subsection we first offer intuition into why that is precisely, and continue with a formal description of the U-CMAB.

The U-CMAB, intuitively. Let $Y \in \{0, 1\}$ for this example. Then, we gain intuition by regarding the bandit problem as a Markov decision process (MDP), an idealised mathematical framework used in *decision modeling* and control to describe the dynamics of an environment [17]. Typically, such an MDP is used in sequential decision modeling and reinforcement learning, whereby applying T to $\mathbf{x}$ presents a subsequent, related context [17]. However, in case of a CMAB the notion of any sequential context is unnecessary and every transitioned state is considered terminal. When terminal nodes are present in the MDP, the framework is termed a *finite* MDP [17, Chapter 3].

The MDP for a CMAB is governed by a transition function, $\mathbb{t}(Y, T, \mathbf{x}) = p(Y|T, \mathbf{x})$, describing the probability of Y after application of T on $\mathbf{x}$. If we Y be binary, we can imagine four different types of individuals:

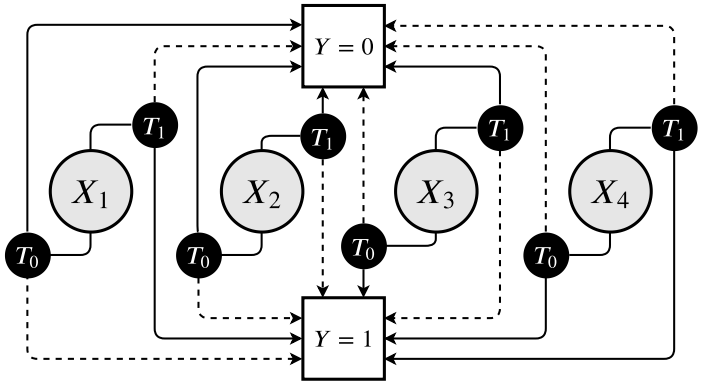


Figure 2. Finite MDP for a CMAB in CATE optimisation. Grey circles denote individuals of type $X_j \subset \mathscr{X}$; squares indicate the response $Y = 1$ or $Y = 0$; black circles represent treatments with T_0 if $T = 0$ and T_1 if $T = 1$, done so for brevity; dashed arrows are used when $\mathbb{t}(Y,T,X_j) = 0$; and full arrows are used when $\mathbb{t}(Y,T,X_j) = 1$.

Persuadables, $X_1 \subset \mathscr{X}$**.** Respond ($Y = 1$) only when treated ($T = 1$),
Lost causes, $X_2 \subset \mathscr{X}$**.** Never respond ($Y = 0$), regardless of treatment,
Sure things, $X_3 \subset \mathscr{X}$**.** Always respond ($Y = 1$), regardless of treatment,
Do-not-disturbs, $X_4 \subset \mathscr{X}$**.** Respond ($Y = 1$) only when untreated ($T = 0$).

We illustrate these types as a schematic of the MDP in Fig 2. Having this categorisation is how we connect the CMAB-litarture, to the CATE literature as this categorisation stems directly from ITEs [11, 71, 72][7].

If $Y = 1$ is the desired outcome, one can deduct from Fig 2, that only individuals from X_1 yield a positive causal relationship between T and Y as applying treatment (i.e., following T_1) to any other type of individual will either: not result in $Y = 1$; or will, regardless of T. As an example, take the individuals in X_4: as both $T = 1$ and $T = 0$ yield a transition probability of $\mathbb{t} = 1$, it does not matter which treatment the agent applies for the individuals to respond ($Y = 1$). Therefore, a causal agent should only apply treatment ($T = 1$) when given an individual from X_1.

When operating on $\hat{r}$ (as in Eq.(2 and Eq.(3)) to differentiate between treatments, an agent would not find an optimum in case of X_4. However, using penalties, ψ_t, we can further differentiate between treatments and incorporate τ. While we illustrate our causal setup in a simplified manner, the binary constraint can be lifted without loss of generality.

The U-CMAB, formally. As noted above, the U-CMAB should mimic the bandit paradigm for learning while optimising for ITE. We argue that it should do this by incorporating τ. This is done through the following by defining the reward as,

$$r_u(T = t, \mathbf{x}) \doteq \mathbb{E}[Y(t) - \psi_t | \mathbf{x}], \tag{6}$$

where ψ_t respects Eq.(5) and optimal treatment selection is based on an unbiased estimate of r_u, ensures that a bandit algorithm optimises for CATE according to the optimal treatment threshold τ, instead of response only.

We motivate Eq.(6) by deriving τ from treatment selection procedures of bandits (as in Eq.(3) but using r_u instead,

$$T^* = \arg\max_i \{r_u(T = i, \mathbf{x})\}.$$

In the binary treatment setting, we can simplify the above as,

$$T^* = \mathbb{I}[r_u(T = 1, \mathbf{x}) > r_u(T = 0, \mathbf{x})],$$

7 Also see Local Average Treatment Effect (LATE), Conditional Average Treatment Effect (CATE), or Complier Average Causal Effect (CACE). Note that the categorisation in [72] and [71] is defined in a noncompliance setting, referring to whether or not the individual accepts treatment. While the motivation of their categorisation might be different, the mathematical machinery is exactly the same.

which can be reformulated using r in Eq.(2) as,

$$T^* = \mathbb{I}[r(T=1,\mathbf{x}) - r(T=0,\mathbf{x}) > \psi_1 - \psi_0],$$

which through linearity of $\mathbb{E}[\cdot]$ equals, and Eq.(1) and Eq.(5),

$$T^* = \mathbb{I}[u(\mathbf{x}) > \tau]. \tag{7}$$

With the above (in Eq.(7)), we achieve $T_b^* = T_u^*$; meaning that any bandit optimising for r_u, rather than r as is standard, will optimise the CATE of a treatment. Note that incorporating costs is a trivial adjustment to bandits, however, the connection between bandits and CATE was previously unexplored.

Consider Algorithm 1, describing a general bandit-algorithm with r_u as an optimisation target. In our experiments, we implemented this algorithm with ε-greedy and LinUCB. We motivate these choices in our experimental setup (Section 5.3) as the choice of bandit algorithm does not impact our formal description of a U-CMAB. In hopes to provide more clarity, we provide our code online at https://github.com/vub-dl/u-cmab.

Algorithm 1: general bandit with $\hat{r}_u$

input : An environment $\mathscr{E}$, some function estimate $\hat{r}_u$, and a function R to translate $Y \to \mathbb{R}$.

while *True* **do**

sample some $\mathbf{x}$ from $\mathscr{E}$;
$T \leftarrow$ `explore(`$\mathbf{x}$`)` or `exploit(`$\mathbf{x}, \hat{r}_u$`)`;
$Y \leftarrow$ `apply(`T`)` on $\mathbf{x}$ in $\mathscr{E}$;
$\hat{r}_u \leftarrow$ `update(`$\mathbf{x}, T, Y$`)`;

5 Experiments

While the instability of CATE models in dynamic environments is touched upon in the literature [4], it is never shown explicitly. As such, we have devoted our experiments section not only in light of evaluating the U-CMAB, but also to an illustration of state-of-the-art CATE methods in a dynamic environment, in Section 5.1.

To evaluate the U-CMAB, we have indicated in Section 3 that a CMAB's objective is two-fold: (i) response prediction; and (ii) optimal treatment selection. As such, like a CMAB, we evaluate the U-CMAB for these two objectives separately, where its objectives are: (i) CATE prediction, instead of response prediction (in Section 5.2); and (ii) optimal treatment selection, where optimality is now defined in terms of causality (in Section 5.3).

Throughout our experiments we compare against the current state-of-the-art, being an uplift random forest (URF) [11, 51]. We optimise every URF using a random search over their hyperparameters.

5.1 CATE optimisation in dynamic environments

We claim that contemporary CATE optimisation models perform poorly in a dynamic setting. Our hypothesis is tested with two state-of-the-art methods: an URF [51]; and an extreme gradient boosting method (XGB) (`XGBRegressor`) with a transformed outcome [73]. For both methods, we have conducted a random search to provide optimal hyperparameters.

Evaluation. We will evaluate the performance of above algorithms on a hold-out test set. Test set performance is then plotted as a cumulative uplift chart [54], as is standard in the CATE optimisation literature [11, 53, 58]. Such a cumulative uplift chart will, for some number of consecutive bins of individuals, calculate the uplift (that is the difference between a treated and controlled outcome) in a hold-out test set,

$$q(b) \doteq \left(\frac{Y_{1,b}}{N_{1,b}} - \frac{Y_{0,b}}{N_{0,b}} \right), \tag{8}$$

where: q is calculated for the first b bins in the given test set; $Y_{i,b}$ are amount of responders after $T = i$ in bin b; and $N_{i,b}$ are the amount of individuals treated with $T = i$ in bin b. For a more comprehensive overview of the Qini metric, we refer to [58], and [74, evaluation].

A model is tasked with sorting the individuals in the test set, such that the individuals with higher estimated CATE are grouped above. As such, a better performing algorithm will have higher q in the first number of bins since the individuals with higher estimated CATE should result in $Y = 1$ only when $T = 1$. Ideally, as is the case with the Gini index, or ROC-curve [75], we aim for the qini-curve to be concentrated in the leftmost corner of its visualisation.

Experimental set-up. In order to properly test CATE models in dynamic environments, we need to employ a simulation [68]. The reason for this is simple; we cannot assess an environment for its dynamics without knowing the underlying distributions [7]. As such, we resort to a similar experimental set-up as was introduced by [76], where the distributions of the environment were changed during data collection, mimicking a sudden drift in the environment [6].

With the simulated environment, we sample twenty thousand random experiments[8]. We sample the treatment randomly, such that it is independent from both the individuals, as well as the potential outcomes. Doing so, makes the causal effect identifiable. During sampling we keep the distributions fixed for the first ten thousand sampled experiments, then change them abruptly before the next ten thousand experiments are sampled, simulating a sudden drift [6, 76].

Next, we train three separate models for each algorithm: one is given the *first* ten thousand data points; one is given the *last* ten thousand data points; and one is given *all* data points, resulting in two static datasets and one dynamic dataset, respectively.

Result. Consider Fig 3, depicting the results for each algorithm given the datasets described above. Both the URF and the XGB method perform well with data coming from a non-dynamic environment. When training is done on data containing a sudden drift, performance drops tremendously, further illustrated by calculating prediction noise. Prediction noise is estimated by randomly shuffling the transformed outcome labels in the dataset and refitting them to the model. Grey areas in Fig 3 indicate the standard deviation of these random fits.

If the model captures an ITE, their cumulative uplift should be far removed from the gray areas surrounding the random line. While this is indeed the case for Fig 3A-3Band Fig 3D-3E, it is not for Fig 3C and Fig 3F. We conclude from this that the model trained on dynamic data fitted noise, while the models trained on static data fitted ITE.

We have repeated this experiment numerous times but only included one outcome per algorithm for brevity. The conclusions drawn from every outcome, however, remained the same. Code for these experiments, as well as, more results can be found in our online code repository (https://github.com/vub-dl/u-cmab).

5.2 Objective (i): CATE prediction

Having confirmed the instability of CATE models (and optimisation) in a dynamic environment, we now turn to evaluating our U-CMAB. That is, a contextual bandit optimising r_u. In this experiment we evaluate our model on prediction, comparing the U-CMAB's outputs with an URF.

Model configuration. To train the U-CMAB we resort to a batch constrained learning method [61, 78], using an artificial neural network (ANN) with: three hidden layers; a rectified linear unit (ReLU) activation: and Adam for gradient optimisation [79], depicted in Fig 4.

As pointed out above, training a model cannot happen in a strictly supervised manner as we only receive one outcome Y_i associated with a treatment $T = i$. As such, for every individual $\mathbf{x}$, the Huber-loss [80] is calculated for the treatment indicated in the dataset, and kept 0 for the counterfactual treatments. This loss is then backpropagated through the entire network, drawing inspiration from training deep q-networks [81] as well as fitted q-iteration [78].

Evaluation. Similarly to our previous experiment, we evaluate both the U-CMAB and the URF using the qini. Note however, that contrasting the previous experiment, we will apply these methods in a multi-treatment setting.

8 There is some work on combining random experiments with observational data [77]. The extension to incorporate these ideas is left for future iterations.

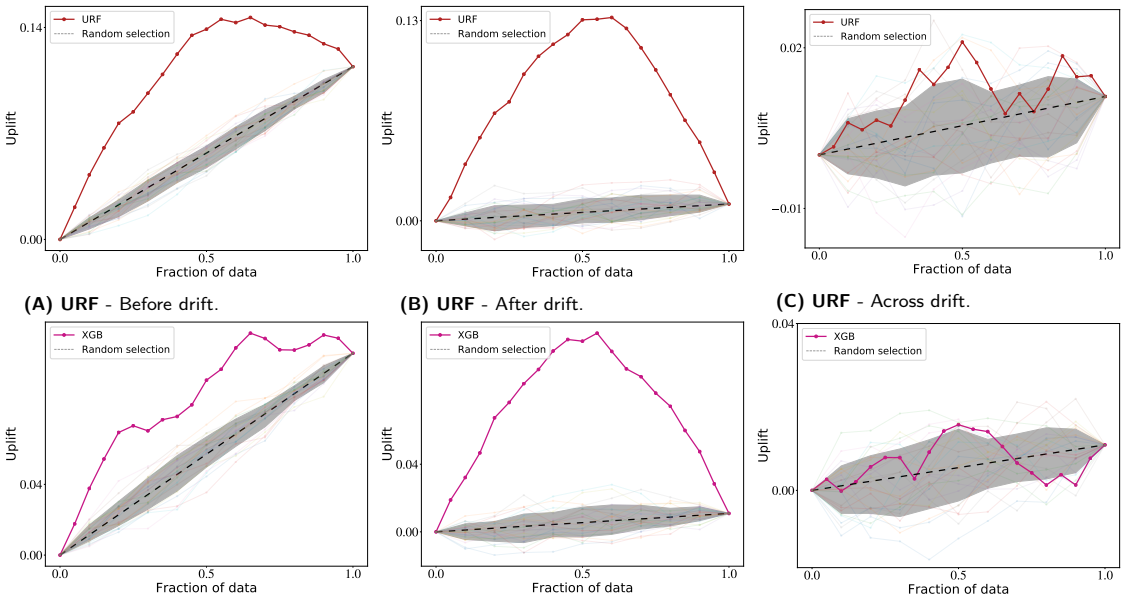


Figure 3. Static CATE models in a drifting environment. qini-charts above illustrate performance by the URF, qini-charts below illustrate performance by XGB. On the x-axis we have the fraction of the data, where each model ranks the data from left-to-right, with highest CATE to lowest ITE. On the y-axis, we calculate the actual CATE (termed uplift) in the data ranked higher than x. Grey areas indicate prediction noise. An ideal model will have a line far above the random line. While models in a static environment seem to accurately predict the CATE (left and middle), when faced with a dynamic environment, performance drops significantly (rightmost).

Experimental set-up. In this experiment we use the Hillstrom-dataset, one of the most used and well-known datasets in CATE optimisation. The Hillstrom-dataset consists of 64k data points in 18 dimensions (after one-hot encoding categorical variables) with 3 treatments (including the control treatment $T = 0$) and a binary response Y. We further motivate this dataset through it being publicly available[9].

Contrasting contemporary methods, the U-CMAB does not require an ensemble of models to handle multiple treatment scenarios— that is, scenarios with $|\mathscr{T}| > 2$. As such, we will compare a batch constrained ANN, with two URFs: URF (T=1) for the first treatment ($T = 1$); and URF (T=2) for the second treatment ($T = 2$). Evaluating these models will thus require two hold-out test sets: a first with $T = 1$ and $T = 0$; and a second with $T = 2$ and $T = 0$. We have conducted this experiment over ten randomly chosen test and training sets, averaged the results, and calculated their standard deviation.

Result. Without a hyperparameter search and a single ANN, the U-CMAB outperforms both URFs— with a random search —in terms of CATE prediction. Moreover, it does so with lesser variance as is shown in Fig 5. This experiment shows that the predictor used by the U-CMAB does indeed predict ITE, thereby confirming objective (i).

While the second treatment in the Hillstrom-dataset ($T = 2$) is notorious for its *unpredictability* [82] — resulting in a seemingly less impressive model — the U-CMAB consistently remains above the random selection line, contrasting URF (T=2). We ascribe this performance to its *single-model* nature: as the U-CMAB shares information across treatments, it could potentially enhance prediction overall, e.g., a healthy patient will still be healthy after both medication ($T = 1$ and $T = 2$) and no medication ($T = 0$).

9 https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

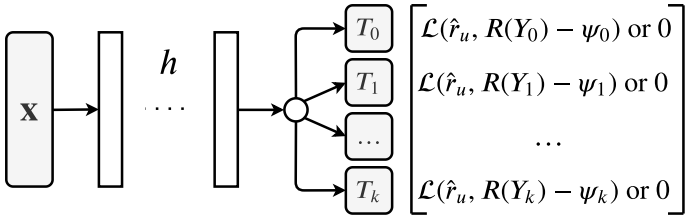


Figure 4. A general ANN architecture for CATE estimation. Having $\mathbf{x}$ as the input; h the hidden layers; $\mathscr{L}$ a loss function; T_i is short for $T = i$; and Y_i is short for Y after $T = i$. Note that in a biased sample setting, we use the flexibility of ANNs to incorporate any balancing terms to counter covariate shift, such as those presented in [10].

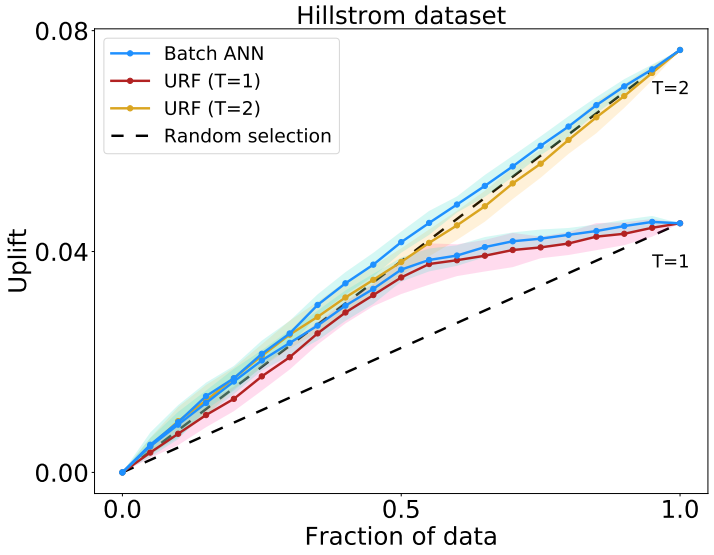


Figure 5. Qini performance on Hillstrom data. Using the Hillstrom dataset (with $k = 10$-fold validation) we compare a *single* batch constrained ANN against *two separate* URFs, where: URF (T=1) was trained for treatment $T = 1$; and URF (T=2) was trained for $T = 2$. We find our model to outperform both URFs consistently.

5.3 Objective (ii): optimal treatment selection

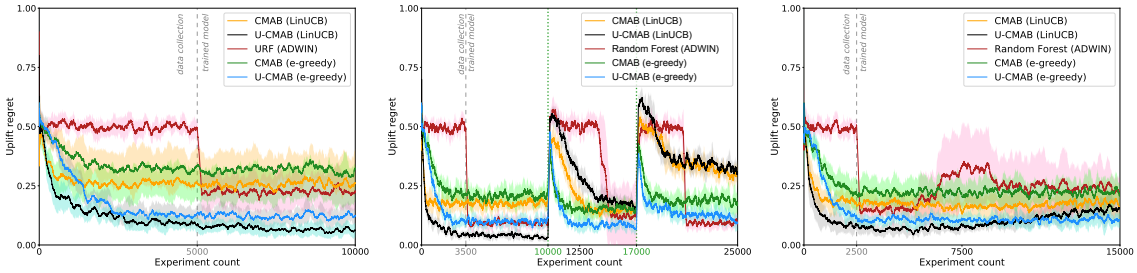
In Section 4.1 we stated that optimality in a bandit setting does not equal optimality in an CATE optimisation setting as, for ITE, optimality was defined in terms of effect rather than response. If the U-CMAB does indeed optimise the ITE, our experiments should indicate that optimal treatment selection for the U-CMAB is the same as in CATE optimisation while following a similar learning paradigm as a CMAB.

In this experiment we test whether the U-CMAB does indeed make optimal, causal decisions. As such, we consider this experiment to be most important to validate our model.

Evaluation. As is customary in bandit literature, we employ a measure of regret to evaluate performance of both algorithms, where algorithmic performance is compared to optimal performance. However, due to the fundamental problem of causal inference, we have to resort to a simulation to calculate this regret as there is no other way to know when a treatment is truly optimal.

We compare both algorithms to the optimal policy, where the optimal policy is one that makes perfectly informed, causal decisions given τ and some environment M. Treatment decisions T_A made by an algorithm A are thus compared to the treatment decisions made by an all knowing policy T^*. As such, we define regret $\mathscr{R}_A(M, \tau)$ as,

$$\mathscr{R}_A(M, \tau) \doteq \mathbb{E}\left[\mathbb{I}\left(T^*(M, \tau) \neq T_A(M, \tau)\right)\right], \tag{9}$$



(A) No drift **(B)** No drift **(C)** No drift

Figure 6. Average performance in simulated environments. We compare our U-CMAB to a CMAB and URF, in $k = 10$ distinct random environments. The grey dashed line indicates the end of the first data gathering period for the URF, yielding a regret of 0.5 as treatments are applied randomly. Dotted lines in Fig 6B indicate a sudden drift. From these experiments, it is clear that across environments, the U-CMAB is performing best.

where every decision $T_A(M, \tau)$ made by A, is compared to the optimal decision $T^*(M, \tau)$. Contrasting the PEHE measure [35], Eq.(9) is based on actual made decisions instead of CATE prediction performance. If every decision is different from the optimal policy, the expectation will be 1, and 0 when they are the same. As such, $\mathscr{R}_A(M, \tau)$—like standard regret —indicates how much A differs from the optimal policy [83].

However, in Section 4.1 we noted that output from an CATE model varies from any bandit. In order to comply with the requirement of decisions as output in Eq.(9), we use Eq.(4) to translate the URF's output to binary decisions.

Experimental set-up. We compare: (i) the U-CMAB (using ε-greedy and LinUCB); (ii) a standard CMAB (using ε-greedy and LinUCB); and (iii) a URF in terms of $\mathscr{R}_A(M, \tau)$. We do this in: a static environment; a suddenly drifting environment as in [76]; and a gradually drifting environment as defined by [6]. Our choice of bandit algorithms is motivated as follows:

ε-greedy - Equipped with a windowed Robins-Monro estimator, ε-greedy is resilient to non-stationary environments, making it a very attractive candidate for our setup. Despite its simplicity, it is reported to perform rather well in a variety of settings [63]. Using Robins-Monro, our bandit slowly *forgets* older samples by placing more emphasis on newer samples.

LinUCB - While LinUCB is less resilient in non-stationary settings, it often forms the basis of contemporary agents. Showing that LinUCB is positively affected by our contribution, could thus provoke further research.

As the URF is a static model, we use the ADaptive WINdow (ADWIN) change detection algorithm [7] to indicate when a drift occurred. In order to comply with the strong ignorability assumption, data is then recollected and the model is retrained as would happen in a naive solution to the problem.

We ran every experiment on ten randomly generated environments and averaged the results. These environments were comprised of: one treatment, $T \in \{0, 1\}$; a linear drift; and a random dimension count between 5 and 8 [68].

Result. The results depicted in Fig 6 clearly indicates a performance increase in both dynamic and static environments.

In Fig 6A we confirm better performance of the causal methods. This is a satisfying result as a regular CMAB does not optimise the ITE, and is thus expected to perform poorly when compared to CATE optimisers such as the U-CMAB and the URF.

Next, we notice poor performance of the URF in a dynamic environment, even when ameliorated with an ADWIN change detection strategy. Take Fig 6B which evaluates performance in a suddenly drifting environment, where every sudden drift is indicated by a dotted (green) line. As, on average, the ADWIN strategy notices a sudden drift in ~ 75 experiments, we confirm this poor performance is not due a malfunctioning change detection, but rather the requirement of a data collection phase. While bandit algorithms automatically comply with the strong ignorability assumption, static methods such as the URF do not, and need new data to enforce it. Not complying to strong ignorability has poor performance as a result, as is shown in our previous experiment.

Similarly in Fig 6C, where performance is measured in a gradually drifting environment. The standard deviation of the URF's performance with ADWIN, indicates that change detection is much more unreliable in such an environment, as was already shown by [7]. While we have experimented with more sensitive change detection, the performance did not increase. Even more so, the performance decreased in the suddenly drifting environment due to over-sensitivity in moments of static distributions. We refer the interested reader to our online code repository and appendix for these experiments.

Also note that, while the LinUCB agents perform very well in a stationary environment (see Fig 6A), they respond less well in non-stationary environments as is indicated in Fig 6B and 6C. This is due to their exploration strategy which is based on an upper confidence bound that is calculated with *every* data point.

In both Fig 6B and 6C, the performance of a standard CMAB is favourable to the URF due to its inherent resilience against a dynamic environment. We deduct from this that using online learners such as a non-stationary CMAB is preferable to CATE optimisers when the environment is dynamic.

6 Conclusion and future work

In this paper we proposed a new optimisation target such that any bandit will optimise the CATE in a dynamic environment. Instead of merely optimising response, the U-CMAB optimises the *effect* it has on response, i.e., only when it can actually cause an individual to respond will the U-CMAB apply treatment.

While our optimisation target involves only a modest practical adjustment through r_u, we believe it to have a greater methodological impact as we link two major fields in machine learning: (i) from a causal viewpoint, we introduced an CATE optimisation method that adapts appropriately to a dynamic environment; (ii) from a bandit viewpoint, we introduced a causal bandit, able to make more intelligent choices when applying treatment. As this method stands on the shoulders of bandit literature, future extensions will scale with innovations in this very active field.

Advantages. We identify three advantages of our model: (i) our model outperforms the state-of-the-art in both static, and dynamic environments; (ii) as the optimised target is defined generally, any bandit algorithm could be transformed in an uplifted contextual multi-armed bandit; (iii) in our comparison to standard bandits, we believe we have contributed to both bandit and causal literature, potentially initiating a broad stream of further research.

Limitations. We identify three limitations: (i) while commonly accepted in CATE estimation and optimisation literature, we still assume strong ignorability, lifting this assumption is currently an active direction of future research; (ii) practitioners might be hesitant to deploy a continuous machine learning task for computational reasons, however, the same holds when implementing any of the naive approaches described in our introduction; (iii) performance is dependent on both the causal nature of the treatment and the performance of the chosen bandit algorithm, two parameters that are impossible to take into account before any model is trained.

Future work. We are interested in extending our method to full reinforcement learning, potentially allowing a reinforcement learning agent to make sequential causal-decisions leading to more efficient use of resources. Efficiently managing resources in function of reward could greatly affect adoption in practical settings. Another line of future research is in a multi-agent setting. We believe that causal reasoning could have great impact in problems such as credit assignment [84], a major area of research in multi-agent reinforcement learning. Furthermore, lifting the strong ignorability assumption through the inclusion of a fully learned causal diagram will not only motivate adoption, but it will also allow for easier domain transfer.

Acknowledgements

JB is funded by the W.D. Armstrong Trust Fund. We wish to thank Vincent Ginis, Ann Nowé, Judea Pearl, and our reviewers for their helpful comments.

References

[1] Vegetabile BG. On the Distinction Between" Conditional Average Treatment Effects"(CATE) and" Individual Treatment Effects"(ITE) Under Ignorability Assumptions. arXiv preprint arXiv:210804939. 2021.

[2] Mueller S, Pearl J. Personalized Decision Making - A Conceptual Introduction. UCLA; 2022. R-513.

[3] Mueller S, Li A, Pearl J. Causes of Effects: Learning individual responses from population data. arXiv preprint arXiv:210413730. 2021.

[4] Fang X. Uplift Modeling for Randomized Experiments and Observational Studies [Ph.D. thesis]. Massachusetts Institute of Technology; 2018.

[5] Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. In: Advances in neural information processing systems; 2015. p. 2503-11.

[6] Tsymbal A. The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin; 2004.

[7] Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM computing surveys (CSUR). 2014;46(4):44.

[8] Rubin DB. Causal Inference Using Potential Outcomes. Journal of the American Statistical A. 2005;100(469):322-31. Available from: https://doi.org/10.1198/016214504000001880.

[9] Pearl J. Causality. Cambridge, UK: Cambridge university press; 2009.

[10] Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017. p. 3076-85.

[11] Devriendt F, Moldovan D, Verbeke W. A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. Big Data. 2018;6(1):13-41. PMID: 29570415. Available from: https://doi.org/10.1089/big.2017.0104.

[12] Gubela RM, Lessmann S, Jaroszewicz S. Response transformation and profit decomposition for revenue uplift modeling. European Journal of Operational Research. 2020;283(2):647-61. Available from: https://www.sciencedirect.com/science/article/pii/S0377221719309415.

[13] Holland PW. Statistics and causal inference. Journal of the American statistical Association. 1986;81(396):945-60.

[14] Bottou L, Peters J, nonero Candela JQ, Charles DX, Chickering DM, Portugaly E, et al. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. Journal of Machine Learning Research. 2013;14:3207-60. Available from: http://jmlr.org/papers/v14/bottou13a.html.

[15] Saffari A, Leistner C, Santner J, Godec M, Bischof H. On-line random forests. In: 2009 ieee 12th international conference on computer vision workshops, iccv workshops. IEEE; 2009. p. 1393-400.

[16] Robbins H. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society. 1952;55:527-35.

[17] Sutton RS, Barto AG. Reinforcement learning: An introduction. 2nd ed. Cambridge, MA, USA: MIT press; 2018.

[18] Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669-88.

[19] Pearl J. Causal inference in statistics: An overview. Statistics surveys. 2009;3:96-146.

[20] Imbens GW. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Journal of Economic Literature. 2020;58(4):1129-79.

[21] Bareinboim E, Forney A, Pearl J. Bandits with Unobserved Confounders: A Causal Approach. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. p. 1342-50. Available from: http://papers.nips.cc/paper/5692-bandits-with-unobserved-confounders-a-causal-approach.pdf.

[22] Lattimore F, Lattimore T, Reid MD. Causal Bandits: Learning Good Interventions via Causal Inference. In: Advances in Neural Information Processing Systems 29. Curran Associates, Inc.; 2016. p. 1181-9. Available from: http://papers.nips.cc/paper/6195-causal-bandits-learning-good-interventions-via-causal-inference.pdf.

[23] Lee S, Bareinboim E. Structural Causal Bandits: Where to Intervene? In: Advances in Neural Information Processing Systems 31. Curran Associates, Inc.; 2018. p. 2569-79. Available from: http://papers.nips.cc/paper/7523-structural-causal-bandits-where-to-intervene.pdf.

[24] Neyman J. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Roczniki Nauk Rolniczych. 1923;10:1-51.

[25] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology. 1974;66(5):688.

[26] Kallus N, Puli AM, Shalit U. Removing hidden confounding by experimental grounding. In: Advances in Neural Information Processing Systems; 2018. p. 10888-97.

[27] Alaa AM, van der Schaar M. Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. Available from: https://proceedings.neurips.cc/paper/2017/file/6a508a60aa3bf9510ea6acb021c94b48-Paper.pdf.

[28] Johansson F, Shalit U, Sontag D. Learning representations for counterfactual inference. In: International conference on machine learning; 2016. p. 3020-9.
[29] Kallus N. DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training. In: III HD, Singh A, editors. Proceedings of the 37th International Conference on Machine Learning. vol. 119 of Proceedings of Machine Learning Research. PMLR; 2020. p. 5067-77. Available from: http://proceedings.mlr.press/v119/kallus20a.html.
[30] Zhang Y, Bellot A, van der Schaar M. Learning Overlapping Representations for the Estimation of Individualized Treatment Effects. In: Chiappa S, Calandra R, editors. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. vol. 108 of Proceedings of Machine Learning Research. PMLR; 2020. p. 1005-14. Available from: http://proceedings.mlr.press/v108/zhang20c.html.
[31] Curth A, van der Schaar M. Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms. In: Banerjee A, Fukumizu K, editors. Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. vol. 130 of Proceedings of Machine Learning Research. PMLR; 2021. p. 1810-8. Available from: http://proceedings.mlr.press/v130/curth21a.html.
[32] Berrevoets J, Jordon J, Bica I, gimson a, van der Schaar M. OrganITE: Optimal transplant donor organ offering using an individual treatment effect. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. vol. 33. Curran Associates, Inc.; 2020. p. 20037-50. Available from: https://proceedings.neurips.cc/paper/2020/file/e7c573c14a09b84f6b7782ce3965f335-Paper.pdf.
[33] Hassanpour N, Greiner R. Learning disentangled representations for counterfactual regression. In: International Conference on Learning Representations; 2019. .
[34] Yoon J, Jordon J, van der Schaar M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In: International Conference on Learning Representations; 2018. .
[35] Hill JL. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics. 2011;20(1):217-40.
[36] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association. 2018;113(523):1228-42.
[37] Yao L, Li S, Li Y, Huai M, Gao J, Zhang A. Representation Learning for Treatment Effect Estimation from Observational Data. In: Advances in Neural Information Processing Systems; 2018. p. 2633-43.
[38] Zhang Y, Berrevoets J, Van Der Schaar M. Identifiable Energy-based Representations: An Application to Estimating Heterogeneous Causal Effects. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2022. p. 4158-77.
[39] Bica I, Alaa AM, Jordon J, van der Schaar M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In: International Conference on Learning Representations; 2020. Available from: https://openreview.net/forum?id=BJg866NFvB.
[40] Lim B, Alaa AM, van der Schaar M. Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.; 2018. Available from: https://proceedings.neurips.cc/paper/2018/file/56e6a93212e4482d99c84a639d254b67-Paper.pdf.
[41] Schulam P, Saria S. Reliable decision support using counterfactual models. In: Advances in Neural Information Processing Systems; 2017. p. 1697-708.
[42] Soleimani H, Subbaswamy A, Saria S. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. In: Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017. AUAI Press; 2017. .
[43] Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. Longitudinal data analysis. 2009;553:599.
[44] Roy J, Lum KJ, Daniels MJ. A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. Biostatistics. 2016;18(1):32-47.
[45] Berrevoets J, Curth A, Bica I, McKinney E, van der Schaar M. Disentangled Counterfactual Recurrent Networks for Treatment Effect Inference over Time. arXiv preprint arXiv:211203811. 2021.
[46] Melnychuk V, Frauen D, Feuerriegel S. Causal Transformer for Estimating Counterfactual Outcomes. arXiv preprint arXiv:220407258. 2022.
[47] Berrevoets J, Alaa A, Qian Z, Jordon J, Gimson AE, Van Der Schaar M. Learning Queueing Policies for Organ Transplantation Allocation using Interpretable Counterfactual Survival Analysis. In: International Conference on Machine Learning. PMLR; 2021. p. 792-802.
[48] Rzepakowski P, Jaroszewicz S. Decision trees for uplift modeling. In: 2010 IEEE International Conference on Data Mining. IEEE; 2010. p. 441-50.
[49] Olaya D, Vásquez J, Maldonado S, Miranda J, Verbeke W. Uplift Modeling for preventing student dropout in higher education. Decision Support Systems. 2020;134:113320.
[50] Rzepakowski P, Jaroszewicz S. Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems. 2012;32(2):303-27.

[51] Guelman L, Guillén M, Pérez-Marín AM. Uplift random forests. Cybernetics and Systems. 2015;46(3-4):230-48.
[52] Sawant N, Namballa CB, Sadagopan N, Nassif H. Contextual Multi-Armed Bandits for Causal Marketing. arXiv preprint arXiv:181001859. 2018.
[53] Diemert Eustache, Betlei Artem, Renaudin C, Massih-Reza A. A Large Scale Benchmark for Uplift Modeling. In: Proceedings of the AdKDD and TargetAd Workshop, KDD, London,United Kingdom, August, 20, 2018. ACM; 2018. .
[54] Kane K, Lo VS, Zheng J. Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. Journal of Marketing Analytics. 2014;2(4):218-38.
[55] Verbeke W, Olaya D, Berrevoets J, Verboven S, Maldonado S. The foundations of cost-sensitive causal classification. arXiv preprint arXiv:200712582. 2020.
[56] Berrevoets J, Imrie F, Kyono T, Jordon J, van der Schaar M. To Impute or not to Impute?–Missing Data in Treatment Effect Estimation. arXiv preprint arXiv:220202096. 2022.
[57] Hazlett C. Estimating Causal Effects of New Treatments Despite Self-Selection: The Case of Experimental Medical Treatments. Journal of Causal Inference. 2019;7(1):20180019. Available from: https://www.degruyter.com/view/journals/jci/7/1/article-20180019.xml.
[58] Gutierrez P, Gérardy JY. Causal Inference and Uplift Modelling: A Review of the Literature. In: International Conference on Predictive Applications and APIs; 2017. p. 1-13.
[59] Auer P, Cesa-Bianchi N, Freund Y, Schapire RE. The nonstochastic multiarmed bandit problem. SIAM journal on computing. 2002;32(1):48-77.
[60] Chapelle O, Li L. An Empirical Evaluation of Thompson Sampling. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, editors. Advances in Neural Information Processing Systems 24. Curran Associates, Inc.; 2011. p. 2249-57. Available from: http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf.
[61] Swaminathan A, Joachims T. Batch learning from logged bandit feedback through counterfactual risk minimization. Journal of Machine Learning Research. 2015;16(1):1731-55.
[62] Li L, Chu W, Langford J, Schapire RE. A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the 19th international conference on World wide web. ACM; 2010. p. 661-70.
[63] Kuleshov V, Precup D. Algorithms for multi-armed bandit problems. arXiv preprint arXiv:14026028. 2014.
[64] Zhou L. A survey on contextual multi-armed bandits. arXiv preprint arXiv:150803326. 2015.
[65] Sinha D, Sankararaman KA, Kazerouni A, Avadhanula V. Multi-Armed Bandits with Cost Subsidy. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2021. p. 3016-24.
[66] Koren T, Livni R, Mansour Y. Bandits with movement costs and adaptive pricing. In: Conference on Learning Theory. PMLR; 2017. p. 1242-68.
[67] Berrevoets J, Verboven S, Verbeke W. Optimising individual-treatment-effect using bandits. arXiv preprint arXiv:191007265. 2019.
[68] Berrevoets J, Verbeke W. Causal Simulations for Uplift Modeling. arXiv preprint arXiv:190200287. 2019.
[69] Hansotia B, Rukstales B. Incremental value modeling. Journal of Interactive Marketing. 2002;16(3):35.
[70] Olaya D, Coussement K, Verbeke W. A survey and benchmarking study of multitreatment uplift modeling. Data Mining and Knowledge Discovery. 2020;34(2):273-308.
[71] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. Journal of the American statistical Association. 1996;91(434):444-55.
[72] Angrist J, Imbens G. Identification and estimation of local average treatment effects. National Bureau of Economic Research Cambridge, Mass., USA; 1995.
[73] Jaskowski M, Jaroszewicz S. Uplift modeling for clinical trial data. In: ICML Workshop on Clinical Data Analysis; 2012. .
[74] Wayfair Tech Blog. Pylift: A Fast Python Package for Uplift Modeling; 2021. Available from: https://github.com/df-foundation/pylift.
[75] Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine learning. 2001;45(2):171-86.
[76] Mellor J, Shapiro J. Thompson Sampling in Switching Environments with Bayesian Online Change Detection. In: Carvalho CM, Ravikumar P, editors. Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics. vol. 31 of Proceedings of Machine Learning Research. Scottsdale, Arizona, USA: PMLR; 2013. p. 442-50. Available from: http://proceedings.mlr.press/v31/mellor13a.html.
[77] Hatt T, Berrevoets J, Curth A, Feuerriegel S, van der Schaar M. Combining observational and randomized data for estimating heterogeneous treatment effects. arXiv preprint arXiv:220212891. 2022.
[78] Ernst D, Geurts P, Wehenkel L. Tree-based batch mode reinforcement learning. Journal of Machine Learning Research. 2005;6(Apr):503-56.
[79] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, Conference Track Proceedings; 2015. Available from: http://arxiv.org/abs/1412.6980.
[80] Murphy KP. 7. In: Dietrich T, editor. Machine Learning: a Probabilistic Perspective. MIT Press; 2012. p. 225-6.

[81] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529.

[82] Radcliffe N. Winning entry - Hillstrom's MineThatData Email Analytics Challange: An Approach Using Uplift Modeling. Stochastic Solutions Ltd.; 2008. Available from: https://pdfs.semanticscholar.org/04e7/5e01a0b4f282f3674cc549d3e9f89fd38c17.pdf.

[83] Wiering M, van Otterlo M. 6.3.2. In: Reinforcement Learning: State-of-the-Art. vol. 12 of Adaption, Learning and Optimization. Springer-Verlag Berlin Heidelberg; 2012. p. 180-2.

[84] Foerster JN, Farquhar G, Afouras T, Nardelli N, Whiteson S. Counterfactual multi-agent policy gradients. In: Thirty-Second AAAI Conference on Artificial Intelligence; 2018. .