# The Supervised Learning Approach To Estimating Heterogeneous Causal Regime Effects

Thai T. Pham*

August, 2016

**Abstract**

We develop a nonparametric framework using supervised learning to estimate heterogeneous treatment regime effects from observational or experimental data. A treatment regime is a set of sequential treatments. The main idea is to transform the unobserved variable measuring a treatment regime effect into a new observed entity through an estimable weight from the data. With the new "transformed" entity, we can estimate causal regime effects with high accuracy using techniques from the machine learning literature. We demonstrate the effectiveness of the proposed method through simulations. Finally, we apply our method to the North Carolina Honors Program data to evaluate the effect of the honors program on students' performances over multiple years. The results show a strong heterogeneity in the students' optimal treatment regimes.

**Keywords:** heterogeneous causal effect, treatment regime, transformed outcome, nonparametric estimation, machine learning, deep reinforcement learning.

**JEL Classification:** C32, C45, C81.

## 1 Introduction

Estimating causal effects of multiple sequential treatments is of growing interest as the situation becomes more widespread. Patients need to make adjustments in medications in

*Graduate School of Business, Stanford University. Email: thaipham@stanford.edu. I am especially indebted to Guido Imbens for guiding and supporting me throughout the project, which include many insightful discussions and invaluable suggestions. I thank Susan Athey, Weixin Chen, Robert Donnelly, Han Hong, Quoc Le, Suraj Malladi, Sheng Qiang, Peter Reiss, and seminar participants at Stanford Graduate School of Business, Facebook's Economics and Computation workshop, and 2016 Asian Meeting of the Econometric Society for various helpful comments. All errors are mine.

multiple periods; after each period, they observe past effects and decide the medication level to take in the next period. Students decide whether to join a specialized mathematics program over a standard one, and they do it each year after observing results in the past. These examples also illustrate the importance of estimating heterogeneous causal effects rather than average causal effects. A sequence of medications may have little effect on average but may be so effective for some patients and ineffective for others (Mega, Sabatine, and Antman [19]). Similarly, the specialized mathematics program may better fit some types of students while hurting others; this fact will be shown in the empirical part.

In this paper, we develop a nonparametric framework using supervised learning to estimate heterogeneous causal effect of a treatment regime from observational or experimental data. We refer to the causal effect in this case the treatment regime effect. This approach is motivated by the Transformed Outcome idea in Athey and Imbens ([1]). The main point is to transform the unobserved variable measuring a treatment regime effect into a new estimable entity from the data. This new entity is related to the original one through the identification results discussed in Section 3, in particular Theorems 3.2 and 3.3. Then, we can use data with the new entity to estimate heterogeneous causal regime effects using supervised learning techniques.

Our proposed method performs well with sequential multi-valued treatments. It also works for both observational and experimental data; this is useful as it is generally hard to set up sequential randomized experiments in reality. The method's advantage comes from the accurate estimation power of the machine learning approach. Unlike most traditional approaches, our method does not rely on parametric form assumptions. So the causal effect estimates it produces are resistant to functional form misspecification.

We present our ideas in the context of an evaluation of the effect of an Honors Program on students' performances when this program is in use for multiple consecutive years. We use data from the North Carolina Honors Program (NCHP). In this data set, students in ninth and tenth grades could take the standard or the honors track in mathematics. Hence, there are four possible sequences of programs each student could follow: standard-standard, standard-honors, honors-standard, and honors-honors. The detailed setting is as follows.

The initial information about students at the end of eighth grade, $X_0$, is given. This information includes the mathematics test score $Y_0$ and other personal and demographic information such as gender, height, and location. The students decide whether to follow the honors program ($W_0 = 1$) or the standard program ($W_0 = 0$) for their ninth grade. At the end of ninth grade, the mathematics test score $Y_1$ is observed. Then the students decide whether to switch or to stay in the current program ($W_1 = 1$ or 0) for their tenth grade. At the end of tenth grade, the mathematics test score $Y_2$ is observed. The object of interest in

this case is the students' test scores at the end of tenth grade, $Y_2$, though it could be $Y_1 + Y_2$ or any functions of $X_0$ and $Y_1, Y_2$.

Individual students may be interested in comparing the object of interest under different treatment regimes so they can decide which program track to follow before ninth grade. School administrators may be more interested in making students choose the appropriate programs in a dynamic manner; that is, they let students choose the honors or standard program in ninth grade based on observed characteristics in eighth grade and let them choose the right program in tenth grade based on observed information in eighth and ninth grades. We refer to the first decision process the *Static Treatment Regime* model and the latter the *Dynamic Treatment Regime* model. In this paper, we discuss both of these settings.

The rest of the paper is organized as follows. Section 2 discusses the *Treatment Regime* model. The focus is on the dynamic treatment regime, considering the static setting a special case. Section 3 derives the identification results for both the static and dynamic treatment regimes. Section 4 explains the general framework of the supervised machine learning approach. Section 5 introduces the testing method based on matching. Section 6 covers the estimation of the model. Section 7 demonstrates the effectiveness of the proposed method through simulations. Section 8 looks at an empirical application using the NCHP data. Section 9 reviews related literature. Section 10 concludes and discusses future work.

# 2 The Treatment Regime Model

Consider a setting with $(T+1)$ stages or periods where $T \geq 1$. The data set is comprised of variables

$$X_0, W_0, Y_1, X_1, W_1, ..., Y_T, X_T, W_T, Y_{T+1}, X_{T+1}.$$

Here, $X_0$ is the initial set of covariates; for $j \in \{1, ..., T+1\}$, $X_j$ is the set of covariates in period $j$ after receiving treatment $W_{j-1}$ but before receiving treatment $W_j$. In case these covariates do not change across periods ($X_j = X_0$ for all $j$), we simply ignore $X_j$'s. For $j \in \{0, 1, ..., T\}$, $W_j$ is the treatment whose value, or treatment level in period $j$, belongs to the set $\mathcal{W}_j$. In this setting, we assume $\mathcal{W}_j$'s are identical and finite. After applying treatment in each period $j \in \{0, 1, ..., T\}$, we obtain the outcome $Y_{j+1}$ in period $(j+1)$. Fitting this setting to the NCHP data, we have $T = 1$ and $X_0, W_0, Y_1, W_1, Y_2$ are as described in the introduction while $X_1 = X_2 = X_0$ and $\mathcal{W}_0 = \mathcal{W}_1 = \{0, 1\}$.

We follow the notations in Orellana, Rotnitzky, and Robins [26] and define

$$W_j = 0 \text{ for } j < 0;\ X_j = 0 \text{ for } j < 0;\ Y_j = 0 \text{ for } j < 1;\ \text{and}$$

$$O_0 = X_0; \text{ and } O_j = (Y_j, X_j) \text{ for } j \in \{1, 2, ..., T+1\}.$$

We also use overbars with a subscript $j$ to denote the present variable at time $j$ and all its past values. For example, $\overline{O}_j = (O_0, O_1, ..., O_j)$. We use notations with no subscript to denote the whole history: $O = \overline{O}_{T+1}$ and $W = \overline{W}_T$.

Now with each possible realization $w = \overline{w}_T = (w_0, w_1, ..., w_T)$ where each $w_j \in \mathcal{W}_j$, we define a vector of potential outcomes:

$$O(w) = O(\overline{w}_T) = (X_0, Y_1(w_0), X_1(w_0), ..., Y_{T+1}(\overline{w}_T), X_{T+1}(\overline{w}_T)) .$$

We can think about these potential outcomes in the same way we think about the traditional potential outcomes (Rosenbaum and Rubin [30], Imbens [14]). When $T = 0$, these two definitions of potential outcomes coincide and we can write the observed outcome

$$Y_1 = W_0 \cdot Y_1(1) + (1 - W_0) \cdot Y_1(0) = Y_1(W_0).$$

Similarly, when $T = 1$ we have

$$\begin{aligned} Y_2 = W_0 W_1 \cdot Y_2(1,1) \; + \; & W_0(1 - W_1) \cdot Y_2(1,0) + \\ & (1 - W_0) W_1 \cdot Y_2(0,1) \; + \; (1 - W_0)(1 - W_1) \cdot Y_2(0,0) = Y_2(W_0, W_1); \text{ and so on.} \end{aligned}$$

Orellana, Rotnitzky, and Robins [26] note that the notation of $O(w)$ implicitly makes a type of Stable Unit Treatment Value Assumption that a unit's potential outcomes are independent of the treatment patterns followed by other units. More concretely, we make the following assumption.

**Assumption 2.1.** (Consistency) *For each $j \in \{1, ..., T+1\}$, we have*

$$\left(Y_j(\overline{W}_{j-1}), X_j(\overline{W}_{j-1})\right) = (Y_j, X_j).$$

In the NCHP setting, there might be interference among students in the same class so that a student' educational track decision can affect other students' outcomes. In this paper, we assume away this possibility.

We proceed to denote by $\mathcal{O} = \{O(w) \mid \text{ each } w_j \in \mathcal{W}_j\}$ the set of all possible vectors of potential outcomes. We also define $d = (d_0, ..., d_T)$, a sequence of decision rules (or treatment regime) where each $d_j$ is a mapping from the space of current and past information $\overline{O}_j$ and past treatments $\overline{W}_{j-1}$ to the space of available treatments $\mathcal{W}_j$:

$$d_j : \overline{O}_j \times \overline{W}_{j-1} \to \mathcal{W}_j.$$

Let $W^d = (W_0^d, W_1^d, ..., W_T^d)$ denote the treatment sequence if the unit had followed the regime $d$. Likewise, $O^d = O(W^d)$ is the corresponding vector of outcomes.

Notice that $(\mathcal{O}, O, W)$ and $(\mathcal{O}, O^d, W^d)$ are random vectors on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with corresponding probability measures named $P$ and $P_d$. By the definition of $\mathcal{O}$, the marginal distribution of $\mathcal{O}$ is the same under $P$ or $P_d$ as it is unaffected by the treatments followed by the unit. The marginal distributions of $(O, W)$ and $(O^d, W^d)$ are different, however.

Specifically, $\mathbb{P}\left(W_j^d = w_j | \overline{O}_j^d = \overline{o}_j, \overline{W}_{j-1}^d = \overline{w}_{j-1}\right) = \mathbf{1}_{\{w_j = d_j(\overline{o}_j, \overline{w}_{j-1})\}}$ is an indicator function while $\mathbb{P}\left(W_j = w_j | \overline{O}_j = \overline{o}_j, \overline{W}_{j-1} = \overline{w}_{j-1}\right)$ (which we denote by $e_{w_j}(\overline{o}_j, \overline{w}_{j-1})$) is the probability that the treatment $w_j$ is taken in period $j$ for the unit with observed past $\overline{O}_j = \overline{o}_j$ and $\overline{W}_{j-1} = \overline{w}_{j-1}$. Note that $e_{w_j}(\overline{o}_j, \overline{w}_{j-1})$ is similar to the (generalized) propensity score defined for the single (multi-valued) treatment case (Imbens [14]). We let $P^m(O)$ and $P_d^m(O^d)$ be the marginal distributions of $O$ and $O^d$. Similar notations can be inferred.

The setting so far is rather flexible; to derive meaningful identification results, further restrictions must be imposed. We make the Sequential Randomization assumption which makes the data look like it is generated from a series of sequential randomized experiments.

**Assumption 2.2.** (Sequential Randomization) *$W_j$ is conditionally independent of $\mathcal{O}$ given $\overline{O}_j$ and $\overline{W}_{j-1}$ for each $j \in \{0, 1, ..., T\}$.*

In the single treatment case, Assumption 2.2 is referred to as unconfoundedness. This assumption states that conditional on past information, a unit's treatment is independent of all potential outcomes of all units. In the NCHP data, this assumption may not completely hold due to possible selection bias problem. However, addressing this problem is out of the scope of this paper.

Finally, we make the Positivity assumption, which is known as overlap in the single treatment case.

**Assumption 2.3.** (Positivity) *For each $j \in \{0, 1, ..., T\}$ and an arbitrary sequence of realized treatments $(W_0, W_1, ..., W_T)$, there exists $\epsilon > 0$ such that*

$$\epsilon < e_{W_j}\left(\overline{O}_j, \overline{W}_{j-1}\right) = \mathbb{P}\left(W_j | \overline{O}_j, \overline{W}_{j-1}\right) < (1 - \epsilon).$$

This assumption states that any possible realization of the treatment regime is followed by at least one unit. For the NCHP data, this assumption should hold due to the heterogeneity in the student population.

# 3 Identification Results

In this section, we discuss the static and dynamic treatment regimes and the corresponding identification results.

## 3.1 The Static Treatment Regime

Suppose we are interested in estimating the effect of applying the regime $d$ rather than another regime $d'$ on a (user-specified) function $u(O)$, conditional on a set of initial covariates $X_0$.[1] In the NCHP data with $T = 1$, we choose $u(O) = Y_2$ or $Y_{T+1}$. In other cases, we may choose $u(O) = \sum_{t=1}^{T+1} Y_t$ or any function of $X_i$'s, $Y_j$'s. In this setting, we do not assume that each $d_j$ is a mapping from $\overline{O}_j \times \overline{W}_{j-1}$ to $\mathcal{W}_j$, but instead a mapping from $X_0$ to $\mathcal{W}_j$. In other words, the whole treatment regime is decided in advance irrelevant of the intermediate outcomes. Assumption 2.2 is therefore modified to fit this setting.

**Assumption 3.1.** *$W_0, ..., W_T$ and $\mathcal{O}$ are independent of one another conditional on $X_0$.*

Also, $e_{W_j}(\overline{O}_j, \overline{W}_{j-1})$ in Assumption 2.3 is replaced by $e_{W_j}(X_0)$. Now, we define

$$V^d(x) = \mathbb{E}\left[u(O^d)\middle| X_0 = x\right].$$

Then the conditional causal effect that we want to estimate is

$$V^d(x) - V^{d'}(x) = \mathbb{E}\left[u(O^d) - u(O^{d'})\middle| X_0 = x\right]. \tag{3.1}$$

In the NCHP data, $d$ and $d'$ map $X_0$ to $\{0,1\}^2$. Though there are uncountably many regimes, their realizations can be one of only four possible value sets: $(1,1)$ corresponding to the honors-honors options in ninth and tenth grades, $(0,0)$ corresponding to the standard-standard track, $(1,0)$ corresponding to the honors-standard track, and $(0,1)$ corresponding to the standard-honors track. Based on this, we can determine the optimal static treatment regime $d^*(x)$ conditional on the covariates $X_0 = x$:

$$d^*(x) = \arg\sup_d V^d(x).$$

[1] Note that $u(O)$ is a function of $O_j = (Y_j, X_j)$ where $Y_j$ and $X_j$ are in turn functions of $\overline{W}_{j-1}$ so to some extent, we can use $u(O, W)$ as in Orellana, Rotnitzky, and Robins [26] instead of $u(O)$. However, writing $u(O, W)$ can be misleading since $u$ cannot be a direct function of $W$. Because otherwise, the results presented in their paper as well as Theorem 3.2 in ours no longer hold. For example, consider $T = 0$ so we have the static case. Theorem 3.2 implies $\mathbb{E}[u(X_0, Y_1) \cdot W_0/e(X_0)|X_0 = x] = \mathbb{E}[u(X_0, Y_1(1))|X_0 = x]$ for any function $u$. This is true. But if $u$ is instead, say $Y_1 W_0$, then $\mathbb{E}[Y_1 W_0 \cdot W_0/e(X_0)|X_0 = x] = \mathbb{E}[Y_1 \cdot W_0/e(X_0)|X_0 = x] = \mathbb{E}[Y_1(1)|X_0 = x]$ while $\mathbb{E}[Y_1(1)W_0|X_0 = x] = \mathbb{E}[Y_1(1)|X_0 = x] \cdot e(x)$. These two values are generally different.

To estimate (3.1) where the outcome of interest $u(O^d) - u(O^{d'})$ is unobserved, our approach is to transform it through a known or estimable weight. To this end, we define

$$s_j^d(x_0) = \prod_{k=0}^{j} \frac{\mathbf{1}_{\{w_k = d_k(x_0)\}}}{e_{d_k(x_0)}(x_0)} \text{ for each } j \in \{0, 1, ..., T\}.$$

Then we have the identification result for the static treatment regime.

**Theorem 3.2.** *Fix an arbitrary regime $d$. Then under Assumptions 2.1, 3.1, 2.3,*

$$\mathbb{E}\left[u(O) \cdot s_T^d(X_0) \middle| X_0\right] = \mathbb{E}\left[u(O^d) \middle| X_0\right].$$

*Moreover, $s_T^d(X_0)$ is the unique function (up to a.e. $P^m(O|X_0)$) that makes this equation hold for all measurable functions $u$.*

**Proof:** The proof is provided in Appendix B.1. □

Theorem 3.2 implies that for two regimes $d$ and $d'$, we have

$$\mathbb{E}\Big[\underbrace{u(O) \cdot \left[s_T^d(X_0) - s_T^{d'}(X_0)\right]}_{\text{observed/estimable, TO}} \Big| X_0\Big] = \mathbb{E}\big[\underbrace{u(O^d) - u(O^{d'})}_{\text{unobserved, PO}} \big| X_0\big]. \tag{3.2}$$

In randomized experiments, $s_T^d(X_0)$ and $s_T^{d'}(X_0)$ are known while in observational studies, they are estimable. Thus, we have transformed the unobserved difference of potential outcomes (PO) $u(O^d) - u(O^{d'})$ to the estimable outcome $u(O) \cdot \left[s_T^d(X_0) - s_T^{d'}(X_0)\right]$ while preserving the conditional expectation. We refer to this estimable outcome as the transformed outcome (TO). We will discuss later how to use TO to estimate (3.1) and determine $d^*(x)$.

## 3.2 The Dynamic Treatment Regime

In our setting in Section 2, however, treatments are chosen sequentially: a unit chooses (or is assigned) the initial treatment based on its initial covariates; then it subsequently chooses (or is assigned) the next treatment based on the initial covariates, the first treatment, and the intermediate covariates and outcome; and so on. This is the dynamic treatment regime.

Even though the causal effects of static treatment regimes can be of interest in many cases, dynamic regimes are often more interesting from a policy perspective. Moreover, the causal effects of some dynamic regimes can contain more information than that produced by the causal effects of all static regimes.

Another concern about the static regime is the inaccuracy of its causal effect estimation when we consider only the initial covariates. This is because when the regime lasts through

many periods we accumulate a lot more information than just the initial covariates. Dynamic treatment regime setting mitigates this problem by using up all available information.

We define

$$\underline{m}^d_{j,j+l}(\overline{o}_{j+l}, \overline{w}_{j+l}) = \prod_{k=j+1}^{j+l} \frac{\mathbf{1}_{\{w_k = d_k(\overline{o}_k, \overline{w}_{k-1})\}}}{e_{d_k}(\overline{o}_k, \overline{w}_{k-1})}$$

for all $j = 0, 1, ..., T$ and all $l > 0$ such that $j + l \leq T$. We also write

$$\overline{W}_j = \overline{d}_j(\overline{O}_j, \overline{W}_{j-1})$$

to mean

$$W_k = d_k(\overline{O}_k, \overline{W}_{k-1}) \text{ for all } k \in \{0, 1, ..., j\}.$$

Similar notations such as $\overline{W}_j = \overline{d}_j(\overline{O}^d_j, \overline{W}_{j-1})$ are inferred in the same way. The main identification result for the static treatment regime is then stated below.

**Theorem 3.3.** *Fix $j \in \{0, 1, ..., T\}$. Fix an arbitrary regime $d$. Then under Assumptions 2.1, 2.2, 2.3,*

$$\begin{aligned}\mathbb{E}\left[u(O) \cdot \underline{m}^d_{j-1,T}(\overline{O}_T, \overline{W}_T) \middle| \overline{O}_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}, \overline{W}_{j-2})\right] \\ = \mathbb{E}\left[u(O^d) \middle| \overline{O}^d_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}^d_{j-1}, \overline{W}_{j-2})\right].\end{aligned}$$

*Moreover, $\underline{m}^d_{j-1,T}(\overline{O}_T, \overline{W}_T)$ is the unique function (up to a.e. $P^m(O|\overline{O}_j, \overline{W}_{j-1})$) that makes this equation hold for all measurable functions $u$.*

**Proof:** The proof is provided in Appendix B.2. □

Theorem 3.3 implies that if we care only about period $j$ onward, then $\overline{O}^d_j = \overline{O}_j$ which in turn implies

$$\mathbb{E}\left[u(O) \cdot \underline{m}^d_{j-1,T}(\overline{O}_T, \overline{W}_T) \middle| \overline{O}_j, \overline{W}_{j-1}\right] = \mathbb{E}\left[u(O^{\underline{d}_{j-1,T}}) \middle| \overline{O}_j, \overline{W}_{j-1}\right],$$

where $\underline{d}_{j-1,T} = (d_j, ..., d_T)$. Now to determine the optimal dynamic treatment regime, we use a sequential approach. In the last period, we have

$$\mathbb{E}\Bigg[\underbrace{u(O) \cdot \left[\frac{\mathbf{1}_{\{W_T = d_T\}}}{e_{d_T}} - \frac{\mathbf{1}_{\{W_T = d'_T\}}}{e_{d'_T}}\right]}_{\text{observed/estimable, TO}} \Bigg| \overline{O}_T, \overline{W}_{T-1}\Bigg] = \mathbb{E}\Big[\underbrace{u(O^{W_T = d_T}) - u(O^{W_T = d'_T})}_{\text{unobserved, PO}} \Big| \overline{O}_T, \overline{W}_{T-1}\Big],$$

where $d_T = d_T(\overline{O}_T, \overline{W}_{T-1})$, $d'_T = d'_T(\overline{O}_T, \overline{W}_{T-1})$, $e_{d_T} = e_{d_T}(\overline{O}_T, \overline{W}_{T-1})$, and $e_{d'_T} = e_{d'_T}(\overline{O}_T, \overline{W}_{T-1})$. This identification coincides with that in Athey and Imbens [1] for the

single treatment case. Based on this, we can determine the optimal treatment rule in the last period $d_T^*(\overline{O}_T, \overline{W}_{T-1})$. We then apply Theorem 3.3 with $j = T-1$ and $d_T = d_T^*$:

$$\mathbb{E}\Bigg[\underbrace{u(O) \cdot \frac{\mathbf{1}_{\{W_T = d_T^*\}}}{e_{d_T^*}} \cdot \left[\frac{\mathbf{1}_{\{W_{T-1} = d_{T-1}\}}}{e_{d_{T-1}}} - \frac{\mathbf{1}_{\{W_{T-1} = d'_{T-1}\}}}{e_{d'_{T-1}}}\right]}_{\text{observed/estimable, } \mathbf{TO}} \Bigg| \overline{O}_{T-1}, \overline{W}_{T-2}\Bigg]$$
$$= \mathbb{E}\Big[\underbrace{u(O^{W_{T-1}=d_{T-1}, W_T = d_T^*}) - u(O^{W_{T-1}=d'_{T-1}, W_T = d_T^*})}_{\text{unobserved, } \mathbf{PO}} \Big| \overline{O}_{T-1}, \overline{W}_{T-2}\Big],$$

where the shorthand notations are defined as above. We use only the units who follow the estimated optimal treatment rule $d_T^*$ for the investigation in this period. Based on this, we can determine the optimal treatment rule $d_{T-1}^*(\overline{O}_{T-1}, \overline{W}_{T-2})$ in this period.

Continuing this process, we obtain a sequence of optimal treatment rules $(d_0^*(X_0)$, ..., $d_T^*(\overline{O}_T, \overline{W}_{T-1}))$. Moreover, results from dynamic programming and reinforcement learning (Bellman [2], Sutton and Barto [36]) guarantee that this sequence is the optimal dynamic treatment regime (Zhao et al. [39]).

We note that similarly to the static case, the TOs in this case are known in randomized experiments and estimable in observational studies. The next sections are devoted to the estimation process using transformed outcomes based on these identification results.

# 4 General Framework of Supervised Machine Learning

Thanks to the identification results in Section 3, we can use (observed) transformed outcomes to estimate heterogeneous causal effects. To be more precise, we want to find a relation between a transformed outcome denoted by $T$ and a set of covariates denoted by $C$. For static treatment regime (see Equation (3.2)), $C = X_0$ and $T = u(O) \cdot \left[s_T^d(X_0) - s_T^{d'}(X_0)\right]$. We can define $C$ and $T$ similarly for dynamic treatment regime.

Now assume that there are $N$ observations $(C_i, T_i)$'s for $i = 1, ..., N$. Econometric models have long been closely attached to linear regressions. Specifically, econometricians usually try to estimate the relation

$$T = g(C; \beta) + \epsilon \text{ where } g(C; \beta) = C\beta \text{ and } \mathbb{E}[\epsilon | C] = 0.$$

Assume each $C_i$ is of dimension $K$ and thus $\beta$ is of dimension $K$ as well. Let $\mathbf{C}$ denote the $N \times K$ matrix obtained from $N$ vectors $C_i$'s, and $\mathbf{T}$ be the $N \times 1$ matrix obtained from $N$ values $T_i$'s. Then an estimator $\widehat{\beta}$ for $\beta$ is determined by minimizing the sum of squared

residuals:

$$\widehat{\beta} = \arg\min_{\beta} ||\mathbf{T} - \mathbf{C}\beta||^2.$$

When the matrix $(\mathbf{C}^T\mathbf{C})$ is nonsingular, then $\widehat{\beta} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{T}$. This model is simple, elegant, and easy to derive. However, it suffers from the accuracy problem. Apparently, there are no hyperparameters[2] to tune and the model complexity is fixed at the lowest level.[3] Although this model can tell us about the causal relation between the covariates and the outcome, the results can be very misleading. A big problem is that the specified model form may totally be incorrect. For example if the actual model form is highly nonlinear, then all derived estimators using the linear models are biased and inconsistent. Lastly, traditional econometricians use all observed data to estimate for the parameters; as a result, they cannot or do not test for the performance of the model and the estimators.

With the arrival of new methodologies in Statistical Learning (SL) and Machine Learning (ML), we have more options over the model choice as well as a more systematic way to test the estimators' performances. With this flexibility property, ML has proved to be superior in terms of estimation accuracy in practice. Interested reader can see empirical comparisons of different SL and ML methods with linear regressions in Caruana and Niculescu-Mizil [4] and Morton, Marzban, Giannoulis, Patel, Aparasu, and Kakadiaris [21].

Generally, the method of SL or ML also specifies a relation model between the outcome of interest $T$ and the covariates $C$ (which they call features):[4]

$$T = h(C;\beta) + \epsilon.$$

Here, $h$ is often a highly nonlinear function and $\epsilon$ is still the error term with $\mathbb{E}[\epsilon|C] = 0$. Up until this point, it looks exactly the same as a nonlinear regression model. The difference is that we allow the model to vary in terms of complexity, and there is usually a hyperparameter $\lambda$ with a regularization term $f(\beta)$ to penalize complex models.[5] So to estimate the model above, we minimize

$$||\mathbf{T} - h(\mathbf{C};\beta)||^2 + \lambda f(\beta),$$

The question now is how to decide on the complexity of the model. Normally, more complex

[2]This term will be introduced shortly below.

[3]We can actually fix the model complexity at a higher level by using nonlinear regression model. This is at the cost of interpretability, however. Specifically, we find the estimator $\widehat{\beta}$ for $\beta$ by minimizing the function $||\mathbf{T} - g(\mathbf{C};\beta)||^2$. Assuming that this function is convex in $\beta$, we can find its minimum by using the first order condition: $g_\beta(\mathbf{C};\beta)^T \cdot (\mathbf{T} - g(\mathbf{C};\beta)) = 0$. We can then solve for the optimal $\widehat{\beta}$ of $\beta$ based on the observed data and the specific form of $g$.

[4]Here, we focus on one type of Machine Learning, the Supervised Learning in which the observed data include both features or covariates and outcomes.

[5]For example, $f(\beta) = \sum_i ||\beta_i||$ or $\sum_i ||\beta_i||^2$ corresponding to what are called Lasso or Ridge regularization.

models work better on the "training set" than the less complex ones. By the training set, we mean the set of data used to estimate the model. But doing well on the training set does not necessarily imply a good performance when the model needs to predict the outcome for a new set of covariates. This is referred to as the overfitting problem. To solve this problem, in the SL and ML literatures, people often partition the set of all observed data into three subsets: the training set, the validation set, and the test set.[6]

With a fixed hyperparameter $\lambda$, the training set is used to choose the optimal model in terms of complexity; this is usually done by using cross-validation. For example, in Multilayer Perceptron in Section 6.2, we use cross-validation to choose the optimal numbers of hidden layers and hidden neurons. Normally in $K$-fold cross-validation (where $K$ is usually 5 or 10), the training set is randomly partitioned into $K$ parts. Each time we use $(K-1)$ parts of them to fit the model and evaluate the loss function (discussed below) using the remaining part. We obtain $K$ values of the loss function and take their average as a criterion for comparison. We choose the complexity corresponding to the model with the lowest value. The validation set is then used to choose the optimal hyperparameter $\lambda^*$, and the test set is used to evaluate the performance of the full model.

In any stage of this process, we use a "loss function" as the objective function we try to minimize. Normally, the loss function is defined similarly to that in traditional econometrics. It is the square root of the average of the squared errors or the residuals. To be more precise, let $\widehat{\beta}$ be the estimator for $\beta$ when $\lambda$ is estimated by $\lambda^*$. The loss function with respect to the validation set of data corresponding to model $h$ is

$$\sqrt{\frac{1}{N_{val}}\sum_{i=1}^{N_{val}}\left(T_i - h(C_i;\widehat{\beta})\right)^2},$$

where the sum goes over all the observations in the validation set. This loss is known as the Root Mean Squared Error (RMSE). Similar loss formulae can be derived for other sets. Certainly depending on specific cases, other forms of loss function can be used.

After we find the best model evaluated on the test set, we can use it to predict the value of the outcome $T^{new}$ corresponding to a new set of covariates $C^{new}$. In other words, a SL or ML model can specify a well-performed prediction relation between $C$ and $T$. The traditional econometric approach can also specify a prediction relation; moreover, it can give an interpretable structural relation between the covariates and the outcome. The latter property is usually missing in a SL or ML model when the form of function $h$ is highly nonlinear. How-

[6] In traditional econometrics, the training set coincides with the set of all observations. As a result, there is no validation set or test set.

ever, empirical studies show that SL and ML methods are superior to econometric approach in terms of accuracy when the relations of interest are unknown. In other words, using SL and ML methods gives more accurate prediction relation at the cost of less interpretable structural relation.

If the prediction returned by the traditional econometric approach is inaccurate, then the implied structural relation would be invalid. Hence in many cases, we should target prediction accuracy. Furthermore, there are many situations in which we can use good prediction approaches to quantify causal effects more accurately; this is exactly a form of causal inference. It is the focus of this paper.

# 5 Testing Methodology - Matching Loss

We recall the identification result for static treatment regime (Equation 3.2) keeping in mind that the dynamic case is similar: $\mathbb{E}[T|C] = \mathbb{E}[Z|C]$, where $Z = u(O^d) - u(O^{d'})$ while $T = u(O)\cdot\left[s_T^d(X_0) - s_T^{d'}(X_0)\right]$ and $C = X_0$. Section 4 shows how to use ML to effectively estimate the relation

$$T = h(C;\beta) + \epsilon,\ \mathbb{E}[\epsilon|C] = 0.$$

We can then recover the relation between $Z$ and $C$: $\mathbb{E}[Z|C] = \mathbb{E}[T|C] = h(C;\beta)$.

There are two issues here, however. First, $T$ can have the same conditional expected value as $Z$ given $C$ but the actual realized values of $T$ and $Z$ may be different; this occurs when $T$ has a large variance. Moreover, even when we choose $T$ as close to $Z$ as possible they are in fact two different quantities. Then the loss function used with $T$ may be inaccurate. Take for example the validation set, the loss functions by using $T_i$'s and $Z_i$'s are

$$\sqrt{\frac{1}{N_{val}}\sum_{i=1}^{N_{val}}(T_i - h(C_i;\beta))^2} \text{ and } \sqrt{\frac{1}{N_{val}}\sum_{i=1}^{N_{val}}(Z_i - h(C_i;\beta))^2},\ \text{respectively.}$$

The first loss function can be calculated from the data while the latter cannot, and these two values may certainly differ. Thus, it is not guaranteed that the use of the loss function with $T_i$'s is satisfactory though it can give indicative results.

To fix this issue, we generalize the matching method used by Rolling and Yang [29] from a single binary treatment case. The idea of their method is to pair each unit in the treatment group with a unit in the control group which has similar covariate values. For each pair, they take the outcome of the treated unit and subtract from it the outcome of the controlled one. They then compare these approximated treatment effects with the estimated treatment effects returned by the considered model. This method is more general than the first one

with verified desirable statistics properties; for example, Rolling and Yang [29] show that under some conditions it is selection consistent.

The disadvantage of this method, however, is its expensive computational time especially when the space of covariates is high dimensional. Another disadvantage is that this method requires many units in each category (that is treated and controlled in the single binary treatment case) to have close matchings. Thus in this paper, we adopt the matching based strategy on the validation and test sets. Specifically, we derive a matching based loss function (from now on, *matching loss*) as a mean of evaluation for these two sets. We use the common loss function (with transformed outcome) for cross-validation on the training set. By doing so, we can obtain good performance while avoiding expensive computations.

We derive the matching loss on the test set. We start with the static treatment regime. Assume we want to compare regime $d$ relative to $d'$. By Theorem 3.2,

$$\mathbb{E}\left[u(O)\cdot\left(s_T^d(X_0)-s_T^{d'}(X_0)\right)\middle|\, X_0\right]=\mathbb{E}\left[u(O^d)-u(O^{d'})\middle|\, X_0\right].$$

Assume that the test sample has $M_d$ units with treatment regime $d$ and $M_{d'}$ units with treatment regime $d'$. We fix an integer $M << \min(M_d, M_{d'})$. Next, we construct $M$ pairs of test units that will be used to approximate unit-level causal effects $\tau = u(O^d) - u(O^{d'})$ to be used in turn to evaluate the estimated conditional treatment regime effect. Specifically, we randomly draw $M$ units from the subsample of units with treatment regime $d$ in the test sample. For each $m \in \{1, 2, ..., M\}$, let $x_0^{d,m}$ and $u(o^{d,m})$ be the set of covariates and the evaluated outcome respectively for the $m^{th}$ unit in this drawn sample.

For each unit $m$ of these $M$ units with treatment regime $d$, we find the closest match $x_0^{d',m}$ to $x_0^{d,m}$ among the units with treatment regime $d'$ in the test sample, in the sense that

$$x_0^{d',m}=\arg\min_{x_0^i\,|\,\text{regime}=d'}||x_0^i-x_0^{d,m}||^2.$$

We now obtain $M$ pairs of covariates $(x_0^{d,m}, x_0^{d',m})$'s with the corresponding evaluated outcomes $(u(o^{d,m}), u(o^{d',m}))$'s. For each pair, we define the approximated causal effect as

$$\widetilde{\tau}_m = u(o^{d,m}) - u(o^{d',m}).$$

Let $\widehat{\tau}$ be the estimator of consideration for $\tau$ which fits the data $x_0$ to the transformed outcome $u(o)\cdot\left(m_T^d(\overline{o}_T,\overline{w}_T)-m_T^{d'}(\overline{o}_T,\overline{w}_T)\right)$. We then define the estimated causal effect in each pair as

$$\widehat{\tau}_m=\frac{1}{2}\left(\widehat{\tau}(x_0^{d,m})+\widehat{\tau}(x_0^{d',m})\right).$$

The matching loss for the test set that we look to minimize is

$$\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widetilde{\tau}_m - \widehat{\tau}_m\right)^2}.$$

Similarly, we can derive the matching loss for the validation set.

Next, we consider the dynamic treatment regime. This case (referred to Theorem 3.3) is in fact almost the same as the case of a single multi-valued treatment so the derivation of the matching loss can be replicated.

# 6 Model Estimation

## 6.1 Estimation of Weight

The transformed outcome is the original observed outcome multiplied by a weight; this weight consists of terms either observable (in randomized experiments) or estimable (in observational studies), where each estimable term is a form of the propensity score:

$$\mathbb{P}\left(W_k = d_k\left(\overline{O}_k, \overline{W}_{k-1}\right)|\overline{O}_k, \overline{W}_{k-1}\right) \text{ for a fixed regime } d.$$

We can estimate each term separately using all the techniques applied to estimating the generalized propensity score for the case with a single multi-valued treatment. Such techniques, for examples, are multinomial logistic regression, generalized additive model, classification tree, random forest, gradient boosting, or multilayer perceptron (See Hastie, Tibshirani, and Friedman [10] for a survey).[7] All these methods are already implemented in the $R$ software.

## 6.2 Estimation of The Treatment Regime Effect

Now we have obtained the transformed outcome values. We can then use Machine Learning to estimate the causal effect of a treatment regime. Though many machine learning techniques would work here (Hastie, Tibshirani, and Friedman [10]), in this paper we use a Deep Learning method called Multilayer Perceptron or Multilayer Neural Network.[8] We note that the method of applying deep learning in the reinforcement learning framework (the dynamic treatment regime setting in the paper) is called Deep Reinforcement Learning. The deep learning method has shown its superiority to other machine learning and traditional

[7] In fact, we can use the deep learning method introduced in Section 6.2 to estimate these terms.

[8] Interested reader is referred to Le [17] for more details regarding this method.

econometric methods. Interested reader can learn about its superiority in LeCun, Bengio, and Hinton [18]. The superiority of the deep reinforcement learning method is provided in Mnih et al. [20].

We now give a general idea of MLP while providing the detailed description of this method in Appendix C. We consider a simple model with one hidden layer. This is in fact a supervised learning method (Section 4) where the relation function $h(X)$ is of the nonlinear form:

$$h(X) = \sum_{j=1}^{K} \alpha_j \sigma(\gamma_j^T X + \theta_j).$$

Here, $\sigma$ is a sigmoid function such as $\sigma(x) = 1/(1 + exp(-x))$ and $(\alpha_j, \gamma_j, \theta_j)$'s and $K$ are parameters that need to be estimated.

Estimating the relation $Y = h(X) + \epsilon$ is not straightforward due to the sophisticated form of $h(\cdot)$. To do it, we use a technique developed in computer science literature called backpropagation (Rumelhart, Hinton, and Williams [31]). The full estimation process is described in detail in Appendix C with the modified testing method in Section 5.

Even though deep learning is superior, the reason it is so is not well understood yet. There is only one main theoretical result regarding the properties of deep learning, independently proved by Cybenko [6] and Hornik, Stinchcombe, and White [13] in 1989. They both show that any multilayer neural network or MLP can approximate any continuous function on any compact subset of a Euclidean space to an arbitrary degree of accuracy. This is known as the universal approximation property. Other than this property, the superiority of deep learning generally and deep reinforcement learning particularly is still a mystery.

# 7 Simulations

In this section, we simulate a two-period treatment regime data set and use it to test the ability of our proposed method in adapting to heterogeneity in the treatment regime effect.

## 7.1 Simulation Setup

The setting is inspired by Wager and Athey [37]. We first simulate the original covariates $X_0 \sim U([0,1]^{10})$. The treatment $W_0 \in \{0,1\}$ is chosen after observing $X_0$. Then we simulate $Y_1 \in \mathbb{R}$ with a standard normal noise (we choose $X_1 = X_0$ so we can ignore $X_1$). Then $W_1$ is chosen in $\{0,1\}$. We simulate the final outcome $Y_2 \in \mathbb{R}$ also with a standard normal noise.

Let the propensity scores be

$$\mathbb{P}(W_0 = 1|X_0) = \mathbb{P}(W_1 = 1|X_0) = \mathbb{P}(W_1 = 1|X_0, W_0, Y_1) = 0.5.$$

In the simulation process, the following true effects are known:

$$\tau_1(X_0) := \mathbb{E}\left[Y_1^{W_0=1} - Y_1^{W_0=0} \,\middle|\, X_0\right] \text{ and}$$

$$\tau_2(X_0, W_0, Y_1) := \mathbb{E}\left[Y_2^{W_1=1} - Y_2^{W_1=0} \,\middle|\, X_0, W_0, Y_1\right].$$

Specifically, denoting by $X_0[i]$ the $i^{th}$ component of $X_0$ we simulate $\tau_1$ and $\tau_2$ to be

$$\tau_1(X_0) = \xi(X_0[1])\xi(X_0[2]), \text{ where } \xi(x) = \frac{2}{1 + e^{-12(x-1/2)}}; \text{ and}$$

$$\tau_2(X_0, W_0, Y_1) = \rho(Y_1)\rho(W_0)\xi(X_0[1]), \text{ where } \rho(x) = 1 + \frac{1}{1 + e^{-20(x-1/3)}}.$$

We keep the main effect in each period at zero:

$$\mathbb{E}\left[Y_1^{W_0=1} + Y_1^{W_0=0} \,\middle|\, X_0\right] = \mathbb{E}\left[Y_2^{W_1=1} + Y_2^{W_1=0} \,\middle|\, X_0, W_0, Y_1\right] = 0.$$

We use $50,000$ observations for training, $5,000$ observations for validation, and $5,000$ for testing. We apply our proposed method to estimate the heterogeneous treatment regime effect using multilayer perceptron (using a simple model with one hidden layer) and compare the result with the baseline in which we use a multivariate linear regression.

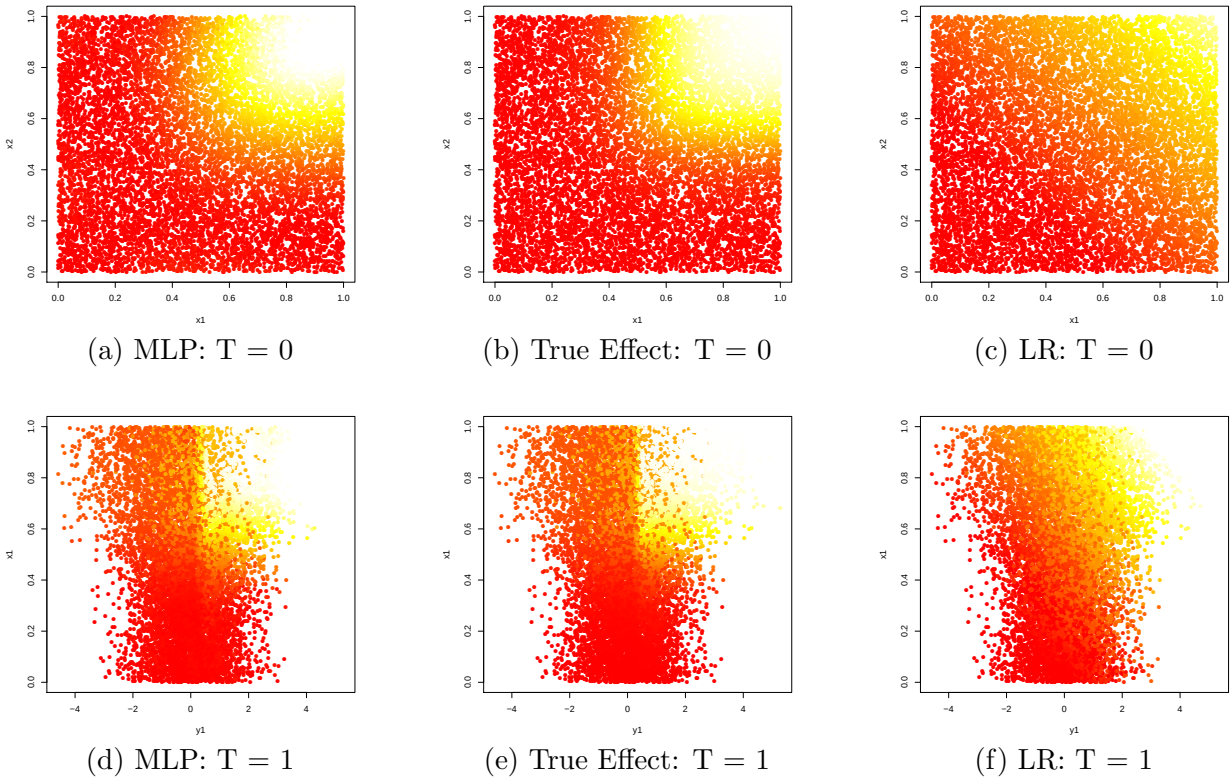
## 7.2 Simulation Results

Estimating the treatment effect for the static and dynamic cases in two periods, we obtain the RMSEs of the estimations (See Table 1). Notice that the standard deviations of the true treatment effects in periods $T = 0$ and $T = 1$ are 1.34 and 2.52 respectively, and those of the transformed outcomes in these periods are 2.41 and 3.21. Thus, Multilayer Perceptron appears to do extremely well and outperform Linear Regression.

Table 1: Performance In Terms of RMSE Using Validation and Test Data.

| Method | Static | Dynamic/T = 0 | Dynamic/T = 1 |
|---|---|---|---|
| Linear Regression | 1.75 | 0.74 | 1.10 |
| Multilayer Perceptron | 1.66 | 0.13 | 0.20 |

Figure 1 visualizes the results.[9]



(a) MLP: T = 0 (b) True Effect: T = 0 (c) LR: T = 0

(d) MLP: T = 1 (e) True Effect: T = 1 (f) LR: T = 1

Figure 1: Heterogeneous Treatment Regime Effect Using Validation and Test Data.
*The two rows correspond to two periods ($T = 0$ and $T = 1$). In each period: the middle heat map visualizes the true treatment effect; the left one is the estimated effect by Multilayer Perceptron; and the right one is the estimated effect by Linear Regression.*

# 8 Empirical Application

We now apply our proposed method to the North Carolina Honors Program data. To recall, this data set consists of two binary treatments $W_0$ and $W_1$, with data appearing in a sequential order: $X_0, W_0, Y_1, W_1, Y_2$.

In the data set, $X_0$ consists of students' scores at the end of eighth grade $Y_0$ and three census dummy variables $d_1, d_2$, and $d_3$. The intermediate outcome $Y_1$ represents students' scores at the end of ninth grade and the outcome of interest $Y_2$ represents students' scores at the end of tenth grade.

There are totally $24,112$ observations in the data set. This data are also pre-scaled so that each non-binary variable has zero mean and unit variance.

[9]We thank Wager and Athey [37] for sharing their visualization code.

## 8.1 Weight Estimation

Since there are only two periods of treatments in this data set, we estimate each term in the weight separately:

$$\mathbb{P}(W_0 = 1|Y_0, d_1, d_2, d_3), \mathbb{P}(W_1 = 1|Y_0, d_1, d_2, d_3), \text{ and } \mathbb{P}(W_1 = 1|Y_0, d_1, d_2, d_3, Y_1, W_0).$$

Here, the first two terms are used for the static treatment regime while the first and the third ones are for the dynamic case. We note that estimating these terms is disguised behind the binary classification process, which is known as probabilistic classification. Thus, to have good estimation requires good classifier.[10] Doing so is tricky because three out of four covariates are binary while the remaining $Y_0$ does not vary much (See Table 2).

Table 2: Summary of Covariate Data

| $d_1$ | $d_2$ | $d_3$ | $W_0$ | Count | mean($Y_0$) | max($Y_0$) | min($Y_0$) |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1539 | -0.21 | 3.67 | -3.02 |
| 0 | 0 | 0 | 1 | 290 | -0.13 | 3.04 | -2.52 |
| 0 | 0 | 1 | 0 | 2107 | -0.35 | 2.92 | -3.02 |
| 0 | 0 | 1 | 1 | 395 | -0.28 | 2.16 | -3.28 |
| 0 | 1 | 0 | 0 | 822 | -0.35 | 3.17 | -3.02 |
| 0 | 1 | 0 | 1 | 181 | -0.49 | 2.66 | -3.15 |
| 0 | 1 | 1 | 0 | 1362 | -0.46 | 3.17 | -3.28 |
| 0 | 1 | 1 | 1 | 274 | -0.41 | 2.28 | -2.52 |
| 1 | 0 | 0 | 0 | 6051 | 0.42 | 4.69 | -2.52 |
| 1 | 0 | 0 | 1 | 1241 | 0.58 | 3.67 | -2.39 |
| 1 | 0 | 1 | 0 | 7222 | 0.22 | 3.67 | -3.02 |
| 1 | 0 | 1 | 1 | 1441 | 0.36 | 3.42 | -2.27 |
| 1 | 1 | 0 | 0 | 415 | 0.24 | 3.42 | -2.27 |
| 1 | 1 | 0 | 1 | 71 | 0.37 | 2.41 | -2.14 |
| 1 | 1 | 1 | 0 | 588 | 0.06 | 3.80 | -2.27 |
| 1 | 1 | 1 | 1 | 113 | 0.13 | 2.16 | -2.39 |

Number of Observations: $24,112$.

With this data set, we suspect if any method would perform much better than random guess since it suffers rather seriously from a form of omitted-variable problem.

[10] Niculescu-Mizil and Caruana [25] examine the relationship between the predictions made by different learning algorithms and true posterior probabilities.

As we can observe, classifying all units to 0 seems like a good choice here as the misclassification rate on the whole data set is only 16.6%. However, in this type of classification the misclassification rate on the treated units is 100% which is very unfavorable. On the other hand, we can classify every unit to 1 to make the misclassification rate on the treated units 0% but the misclassification rate on the untreated or controlled units is then 100% and the overall misclassification rate is 83.4% which is terrible. A reasonable criterion for defining a good classifier would be to make the misclassified treated unit rate less than some *threshold* while making the overall misclassification rate as low as possible. The misclassification rates are calculated using validation set, and the threshold is chosen to balance out these rates.[11]

Because no method is superior than others in this case, we use the popular random forest method to classify on $W_0$ and $W_1$ using the specified criterion above.[12] Note that, we divide the data set into the training, validation, and test sets as described in Section 4. The best parameters are chosen using the above criterion evaluated on the validation set. The threshold mentioned above is chosen to be 0.45 for classifying on $W_0$, 0.38 for classifying on $W_1$ in the static case, and 0.34 for classifying on $W_1$ in the dynamic case. As we have more and better covariates to classify for $W_1$ in the dynamic setting, we would expect better classification result on this. The results are summarized in Table 3.

Table 3: Random Forest For Classifying On $W_0$ and $W_1$

| Treatment | Misclassification Rate | Training Set | Validation Set | Test Set |
|---|---|---|---|---|
| $W_0$ | On Treated Units | 0.41 | 0.45 | 0.51 |
| | Overall | 0.48 | 0.49 | 0.50 |
| $W_1$ (Static) | On Treated Units | 0.38 | 0.38 | 0.42 |
| | Overall | 0.40 | 0.41 | 0.41 |
| $W_1$ (Dynamic) | On Treated Units | 0.32 | 0.33 | 0.35 |
| | Overall | 0.32 | 0.33 | 0.33 |

Next, we discuss the estimation of the full model.

## 8.2 Model Estimation

Using the estimated weight, we obtain the transformed outcome and proceed to the model estimation stage.

[11] Alternative evaluation metrics are Precision and Recall, together with the $F_1$ score.

[12] In fact, Niculescu-Mizil and Caruana [25] find that random forest predicts well calibrated probabilities.

### 8.2.1 Static Treatment Regime

We first consider the static treatment regime (referred to Theorem 3.2). As we are interested in the outcome in the final year only, we set $u(O) = u(\overline{O}_2) = Y_2$. Assume we want to estimate the treatment regime effect of the regime $d = (1, 1)$ relative to the regime $d' = (0, 0)$. The transformed outcome is then

$$Y_{STR}^{TO} = Y_2 \cdot \left[ \frac{W_0 W_1}{\mathbb{P}(W_0 = 1|X_0)\mathbb{P}(W_1 = 1|X_0)} - \frac{(1 - W_0)(1 - W_1)}{\mathbb{P}(W_0 = 0|X_0)\mathbb{P}(W_1 = 0|X_0)} \right]$$

given the covariates $X_0$.

Now we use our proposed method of multilayer perceptron (MLP) to estimate causal effects. We compare this method to linear regression (LR) as a baseline and another popular ML technique named gradient boosting (GB). The results are reported in Table 4.

Table 4: Full Model Estimation - STR

| Method | Validation Matching Loss | Test Matching Loss |
|---|---|---|
| LR | 11.14 | 9.77 |
| GB | 5.03 | 4.89 |
| MLP | 3.20 | 3.27 |

There are three remarks here. First, we use one hidden layer for the MLP model. Second, for the LR method there is no hyperparameter so we use only the training data to fit the model. Third, as described in Section 5 we use the usual RMSE to fit the model on the training set and use matching loss on the validation and test sets. The number of matching loss observations $M$ is chosen to be 300 for both validation and test sets.

### 8.2.2 Dynamic Treatment Regime

Now, we consider the dynamic treatment regime (referred to Theorem 3.3). We first define the transformed outcome in the second (also last) period:

$$Y_{2,DTR}^{TO} = Y_2 \cdot \left[ \frac{W_1}{\mathbb{P}(W_1 = 1|X_0, Y_1, W_0)} - \frac{1 - W_1}{\mathbb{P}(W_1 = 0|X_0, Y_1, W_0)} \right]$$

given the covariates $(X_0, Y_1, W_0)$. Based on this, we determine the optimal treatment rule in this period:

$$d_1^* : (X_0, Y_1, W_0) \to \{0, 1\}.$$

Then we can define the transformed outcome in the first period:

$$Y_{1,DTR}^{TO} = Y_2 \cdot \frac{\mathbf{1}_{\{W_1 = d_1^*\}}}{\mathbb{P}(W_1 = d_1^* | X_0, Y_1, W_0)} \cdot \left[ \frac{W_0}{\mathbb{P}(W_0 = 1 | X_0)} - \frac{1 - W_0}{\mathbb{P}(W_0 = 0 | X_0)} \right]$$

given covariates $X_0$.

In this setting, we would be more interested in the treatment effect of $W_1$ in the second period; and of $W_0$ on the first period, given that $W_1 = d_1^*$.

We run MLP, GB, and LR to estimate the treatment effect of $W_1$ on the second period and report the results in Table 5. Similarly to the STR case, we also use matching loss on validation and test sets and $M = 300$.

Table 5: Full Model Estimation - DTR Period 2 ($T = 1$)

| Method | Validation Matching Loss | Test Matching Loss |
|---|---|---|
| LR | 1.28 | 1.29 |
| GB | 0.94 | 1.01 |
| MLP | 0.94 | 1.01 |

We obtain that $d_1^*(X_0, Y_1, W_0) \equiv 0$. We then estimate the treatment effect of $W_0$ in the first period given that $W_1$ is chosen optimally. Thus, we keep only the observations corresponding to which $W_1 = 0$; there are $16,891$ such observations. The results for the treatment effect in this period are summarized in Table 6. As the size of the data set is smaller, we use $M = 100$ for the matching loss on both validation and test sets.

Table 6: Full Model Estimation - DTR Period 1 ($T = 0$)

| Method | Validation Matching Loss | Test Matching Loss |
|---|---|---|
| LR | 3.29 | 3.45 |
| GB | 1.51 | 1.63 |
| MLP | 1.14 | 1.60 |

The advantage of MLP over GB and certainly over LR lies in its flexibility; we can modify MLP in many ways to fit different tasks. GB is not so restricted, though is less flexible than MLP. Regarding LR, there is no hyperparameter so so the estimated results are fixed across various tasks. As we can observe, the MLP performance is the best on the matching loss

on both validation and test sets for both static and dynamic treatment regimes. Moreover, most of the time MLP also outperforms the other two methods in terms of RMSE.

On the other hand, all three methods do much better in DTR than in STR. This shows the dynamic nature of the data, and that students actually choose their programs sequentially over multiple years. Notice that the standard deviations of the pair (transformed outcome, "true effect") in period $T = 1$ (true effect is computed using sampled matching outcomes on $Y_2$) are $(4.06, 0.85)$ and $(4.03, 0.92)$ for the validation and test sets, respectively; these pairs of values in period $T = 0$ are $(6.94, 0.84)$ and $(7.45, 0.98)$, respectively. Hence, the matching loss values in DTR case are reasonable since they account for the bias of the estimated values, the variance of these values, and the discrepancy in the matching pairs however we pair. The results are worse than that in simulations because of the unobserved heterogeneity. This creates the omitted-variable bias that makes the matching loss much larger (recall that the estimations of propensity scores are highly inaccurate).

Now for each old or new student, we obtain an estimated value for the effect of joining the honors program rather than the standard one in each period or through the whole track. The standard error estimation of this value as well as hypothesis testing will be discussed in Section 10 and Appendix D.

## 8.3 Heterogeneous Optimal Regime

In this section, we investigate the optimal regime and its effects when we take into account individual characteristics. We use MLP to estimate these effects for both static and dynamic treatment regimes.

### 8.3.1 Static Treatment Regime

We first consider the static treatment regime. Because there are two treatment periods, we need to compare four sequences: $(W_0, W_1) = (1, 1), (0, 0), (1, 0), (0, 1)$. Above, we have estimated the heterogeneous effect of $(1, 1)$ relative to $(0, 0)$. We need to investigate two more relations: between $(1, 1)$ and $(1, 0)$, and between $(1, 1)$ and $(0, 1)$.

The Average Treatment Effect (ATE) of $(1, 1)$ relative to $(0, 0)$ is $-4.67$; of $(1, 1)$ relative to $(1, 0)$ is $-0.41$; and of $(1, 1)$ relative to $(0, 1)$ is $-0.41$. Hence if the population is homogeneous, then we would assign all of $24,112$ students to the regime $(0, 0)$ or the standard - standard track. However, the population's heterogeneity implies that different groups would best fit different regimes. If we use the optimal rule for each student, then 653 students would have chosen regime $(1, 1)$; $9,411$ students would have chosen regime $(1, 0)$; $3,256$ would have chosen regime $(0, 1)$; and only $10,792$ students would have chosen regime $(0, 0)$. Figures 2

and 3 visualize this heterogeneity.

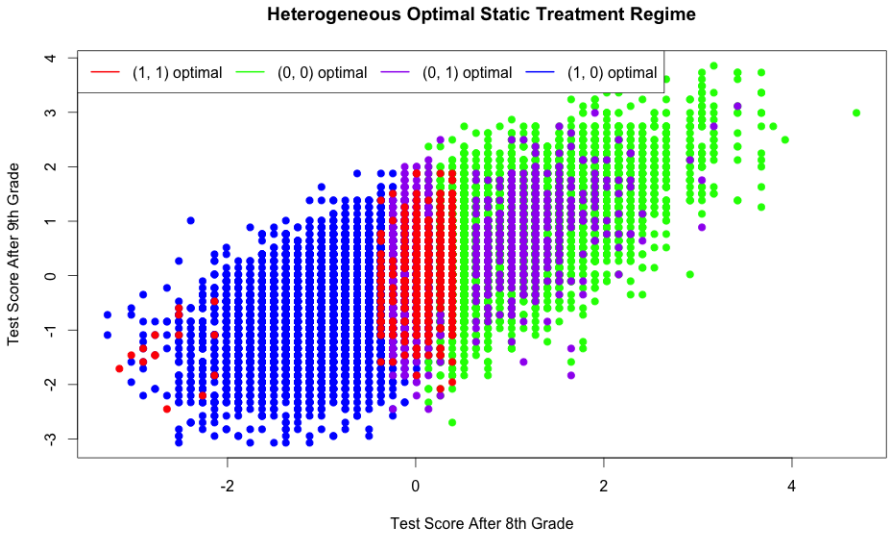



Figure 2: Heterogeneous Optimal STR Over Test Scores After $8^{th}$ and $9^{th}$ Grades.

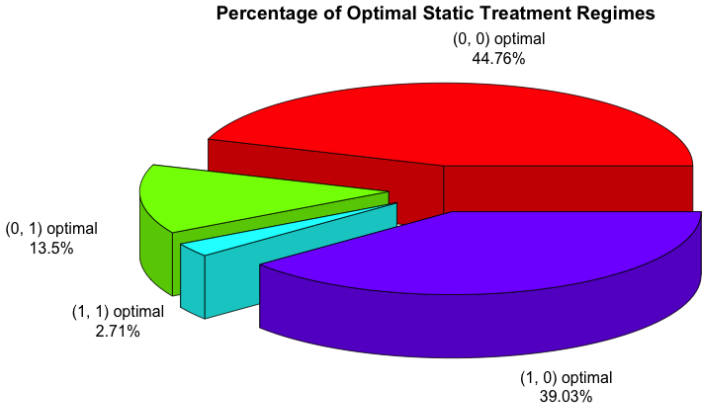



Figure 3: Heterogeneous Optimal STR - Distribution of Regime Types.

We can also explicitly estimate the gain of the heterogeneous optimal assignment over assigning all students to regime $(0, 0)$:

$$\sum_{(1,1) \text{ opt}} \left[\widehat{Y}_2(1,1) - \widehat{Y}_2(0,0)\right] + \sum_{(1,0) \text{ opt}} \left[\widehat{Y}_2(1,0) - \widehat{Y}_2(0,0)\right] + \sum_{(0,1) \text{ opt}} \left[\widehat{Y}_2(0,1) - \widehat{Y}_2(0,0)\right].$$

Dividing this value by the number of students who should not have followed $(0, 0)$ track

$(13, 320)$ gives us the estimated gain on average for each of these students of 0.91. This value is significantly large as the scaled outcome $Y_2$ over the whole data set has zero mean with $\min(Y_2) = -4.06$ and $\max(Y_2) = 3.66$.

### 8.3.2 Dynamic Treatment Regime

Now, we consider the dynamic treatment regime. As mentioned earlier, we consider only students who had followed their predicted optimal treatments in the second period ($T = 1$); there are $16,891$ of them. Consider these individuals in the first period. If we assume the population is homogeneous, then since the average treatment effect of $W_0 = 1$ relative to $W_0 = 0$ is $-1.27 < 0$ we would assign all these students to the standard program in period $T = 0$. However again, the heterogeneity of the population implies that some students would prefer the honors program instead. Specifically $6,183$ among these $16,891$ students prefer the honors program in $T = 0$ while $10,708$ prefer the standard one.

If we follow the rule suggested by the MLP method in $T = 0$, then the average estimated gain for each of the students who should have followed $W_0 = 1$ over assigning all students to $W_0 = 0$ is

$$\frac{\sum_{W_0=1 \text{ optimal}} \left[\widehat{Y}_2^{W_0=1} - \widehat{Y}_2^{W_0=0}\right]}{\#\text{ obs used in } T = 0 \text{ s.t. } W_0 = 1 \text{ opt}} = 0.74.$$

Again, this gain is significant as $Y_2$ has zero mean and is in $[-4.06, 3.66]$.

# 9 Related Literature

There have been several works in applying supervised machine learning techniques to the problem of estimating heterogeneous treatment effects when there is a single treatment. Some of those works are Athey and Imbens [1], Foster, Taylor and Ruberg [9], Su, Tsai, Wang, Nickerson, and Li [35], and Dudik, Langford, and Li [7] in which all of them use some forms of tree-based methods.

Our work is similar in motivation to this line of work in that we transform the problem into the form of a supervised learning problem through the usage of a weight. However, our work differs in that it focuses on treatment regimes instead of settings with single treatments.

Our approach is most similar to Orellana, Rotnitzky, and Robins [26] and Robins, Hernan, and Brumback [28] in our use of a weight in the transformed outcome, which they refer to as the inverse probability weight. However, while we focus on the causal regime effects they focus on determining the optimal regime. Moreover in both of these papers, they need to make specific parametric form assumptions relating the outcome with treatment variables

while our method is nonparametric. Their approaches are also different from our supervised learning approach in that they use some form of approximate dynamic programming techniques through $Q$-functions.

Also using inverse probability weighting technique, Zhao, Zheng, Laber, and Kosorok [39] look at finding the optimal treatment regime. Their approach differs from previous work in that they transform the original problem into a weighted SVM one and solve a convex optimization problem. This way, they estimate the optimal treatment regime rather than attempting to back it out of conditional expectations. While this formulation is interesting and novel, generalizing it to the case when treatment variables are multi-valued instead of binary can be very complicated. Also, it can give only the estimated optimal choice of the treatment regime without giving any quantitative estimate for the causal effect. This is problematic when the decision rule is at the threshold of two different treatment levels and when no validated way is used to test for the correctness of the rule. Most importantly, their method only works for data generated from SMART designs (Murphy [22]) but not for observational data.

Another approach to estimating the optimal treatment regime is $A$-learning in Blatt, Murphy, and Zhu [3]. This method models the temporal relations between past information and future outcomes like other approximate dynamic programming methods such as $Q$-learning. The difference is that instead of using $Q$-functions, they use regret functions. This method also requires some parametric form assumptions. A survey for the $Q$- and $A$-learning can be seen in Schulte, Tsiatis, Laber, and Davidian [32]. A general survey of dynamic programming approaches to estimating the optimal treatment regime can be seen in Chakraborty and Moodie [5].

# 10 Conclusions

In this paper, we introduced a nonparametric framework using supervised learning to estimate heterogeneous causal regime effects. Though this framework can be used equally well for experimental data, we mainly focus on its applications in observational data.

As treatment regimes are ubiquitous in many scenarios, especially education and medicine, effective methods to estimate for their causal effects are always in demand. Our proposed approach is easy to understand compared to alternative methods in the treatment regime area. We utilize the best techniques in machine learning literature, which is mostly used for the prediction purpose to quantify the causal effects of treatment sequences. This opens up the possibility of the use of prediction methods in studying complicated causal inferences when the economic models are set up in the right way. Also, we join a mini-trend

of applying machine learning techniques in economics and social sciences in general.

On the other hand, our method offers a heterogeneous view over effects of treatment regimes; this is particularly important when some subpopulations have much higher than average treatment regime effects while some others have much lower than average values.

Researchers can use our method in the context where the treatments are multi-valued, and especially when the covariate space is large or when the number of periods in which treatments are available is large, or both.

On the other hand, there are several directions the future work can focus on based on our work in this paper. One could try to establish asymptotic results so that we can do formal hypothesis testing. In fact, we can follow White [38] to show the consistency and asymptotic normality of our proposed estimators. However, this is true at the cost of eliminating the regularization term. If we discard this term, then we are back to the overfitting problem that we want to fix at the first place.

When this term is present, the estimators in our machine learning model are unfortunately biased and likely inconsistent (See Appendix D). If so, even when asymptotical normality holds we cannot do any valid hypothesis testing. In Appendix D, we provide a reasonable estimator for the true variance of the treatment regime effect using a matching based kernel method. With regularization, this estimator cannot easily render a formal inferential result but it is certainly indicative. Hence one research direction could be to look for conditions under which we still obtain consistency and asymptotic normality even when a specific form of regularization is present.

Future researchers can also experiment with a deeper network. In this paper, we use only a single hidden layer though two or three or even more hidden layers may work better on certain data sets. Another direction for future research could be empirically testing the performance of the matching loss in different data sets especially thoroughly simulated ones. One can also combine the usual RMSE with the matching loss in the addition manner, and use this loss function instead of the matching one. Experimenting with this loss function can be of interest.

# References


[1] Athey, S. and G. Imbens, (2015), "Recursive Partitioning for Heterogeneous Causal Effects," `http://arxiv.org/pdf/1504.01132v3.pdf`

[2] Bellman, R., (1957), "Dynamic Programming," *Princeton, NJ: Princeton University Press.*

[3] Blatt, D., S. A. Murphy, and J. Zhu, (2004) "A-learning for approximate planning," *Technical Report* 04-63, The Methodology Center, Pennsylvania State University.

[4] Caruana, R. and A. Niculescu-Mizil, (2006), "An Empirical Comparison of Supervised Learning Algorithms," *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA.

[5] Chakraborty, B. and E. E. M. Moodie, (2013), "Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine," *Springer (Statistics for Biology and Health series)*.

[6] Cybenko, G., (1989), "Approximation by Superpositions of a Sigmoidal Function," *Math. Control Signals Systems*, 2: 303-314.

[7] Dudik, M., J. Langford, and L. Li, (2011), "Doubly Robust Policy Evaluation and Learning," *Proceedings of the 28th International Conference on Machine Learning* (ICML-11).

[8] Fan, J. and Yao, Q., (1998), "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, 85, 645-660.

[9] Foster, J., J. Taylor, and S. Ruberg, (2010), "Subgroup Identification from Randomized Clinical Data," *Statistidsdcs in Medicine*, 30, 2867-2880.

[10] Hastie, T., R. Tibshirani, and J. Friedman, (2011), "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Second Edition, *Springer*.

[11] Hochreiter, S. and J. Schmidhuber, (1997), "Long short-term memory," *Neural Computation*, 9(8):1735-1780.

[12] Holland, P., (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association*, 81, 945-970.

[13] Hornik, K., M. Stinchcombe, and H. White, (1989), "Multilayer Feedforward Networkds Are Universal Approximators," *Neural Networks*, Vol. 2, pp. 359-366.

[14] Imbens, G., (2000), "The Role of The Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 3, pp. 706-710.

[15] Imbens, G., and D. Rubin, (2015), "Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction," *Cambridge University Press*.

[16] Knight, K. and W. Fu, (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, Vol. 28, No. 5, 1356 - 1378.

[17] Le, Q. V., (2015), "A Tutorial on Deep Learning - Part 1: Nonlinear Classifiers and The Backpropagation Algorithm," `http://robotics.stanford.edu/~quocle/tutorial1.pdf`

[18] LeCun, Y., Y. Bengio, and G. Hinton, (2015), "Deep Learning," *Nature* 521, 436-444 (28 May).

[19] Mega, J. L. , M. S. Sabatine, and E. M. Antman, (2014), "Population and Personalized Medicine in the Modern Era," *The Journal of the American Medical Association*, 312(19) : 1969-1970.

[20] Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, (2015), "Human-level control through deep reinforcement learning," *Nature* 518, 529-533 (26 February).

[21] Morton, A., E. Marzban, G. Giannoulis, A. Patel, R. Aparasu, and I. A. Kakadiaris, (2014), "A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients," *13th International Conference on Machine Learning and Applications*.

[22] Murphy, S. A., (2005), "An Experimental Design for the Development of Adaptive Treatment Strategies," Statistics in Medicine, 24, 1455-1481.

[23] Murphy, S. A., M. J. Van Der Laan, and J. M. Robins, (2001), "Marginal Mean Models For Dynamic Regimes," *Journal of The American Statistical Association*, Dec, Vol. 96, No. 456, Theory and Methods.

[24] Newey, W. K., (1994), "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, Vol. 10, No. 2, Jun, pp. 233-253.

[25] Niculescu-Mizil, A. and R. Caruana, (2005), "Predicting Good Probabilities With Supervised Learning," *Proceedings of the 22nd International Conference on Machine Learning* (ICML).

[26] Orellana, L., A. Rotnitzky, and J. M. Robins (2010), "Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes," *The International Journal of Biostatistics*, 6(2): Article 8.

[27] Robins, J. M., (1999), "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference," *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, The IMA Volumes in Mathematics and its Applications Volume 116, 2000, pp. 95 - 133.

[28] Robins J. M., M. A. Hernan, and B. Brumback, (2000), "Marginal Structural Models and Causal Inference In Epidemiology," *Epidemiology*, Sep, 11(5): 550 - 60.

[29] Rolling, C. A. and Y. Yang, (2014), "Model selection for estimating treatment effects," *Journal of the Royal Statistical Society, Series B*, Volume 76, Issue 4, pages 749-769.

[30] Rosenbaum and Rubin, (1983), "The Central Role of The Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70 (1): 41-55.

[31] Rumelhart, D. E., G. E. Hinton, and R. J. Williams, (1986), "Learning representations by backpropagating errors," *Nature*, 323 (6088): 533-536.

[32] Schulte, P. J., A. A. Tsiatis, E. B. Laber, and M. Davidian, (2014), "Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Statistical Science*, November, 29(4): 640-661.

[33] Silverman, B.W, (1986), "Density Estimation for Statistics and Data Analysis," *Chapman and Hall*.

[34] Sonoda, S., N. Murata, (2015), "Neural Network with Unbounded Activations is Universal Approximator," *Working Paper*, arXiv:1505.03654.

[35] Su, X., C. Tsai, H. Wang, D. Nickerson, and B. Li, (2009), "Subgroup Analysis via Recursive Partitioning," *Journal of Machine Learning Research*, 10, 141-158.

[36] Sutton, R.S., and A. G. Barto, (1998), "Reinforcement Learning I: Introduction," *Cambridge, MA: MIT Press*.

[37] Wager, S. and S. Athey, (2015), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," Stanford University, Working Paper.

[38] White, H., (1989), "Learning in Artificial Neural Networks: A Statistical Perspective," *Neural Computation* 1, 425-464.

[39] Zhao, Y. Q., D. Zheng, E. B. Laber, and M. R. Kosorok, (2015) "New Statistical Learning Methods For Estimating Optimal Dynamic Treatment Regimes," *Journal of The American Statistical Association*, June, Vol. 110, No. 510, Theory and Methods.

# Appendix A

In this Appendix, we review the Radon-Nikodym theorem and related results which are used in the paper. Throughout this part, we assume that $(\Omega, \mathcal{F})$ is an arbitrary measurable space. We start with the notions of "absolutely continuous" and "$\sigma$-finite" measures.

**Definition A.1.** *Let $\mu$ and $\nu$ be two measures on $(\Omega, \mathcal{F})$. We say that $\nu$ is* **absolutely continuous** *with respect to $\mu$, which we write $\nu \ll \mu$, if and only if for all $E \in \mathcal{F}$, $\mu(E) = 0$ implies $\nu(E) = 0$.*

**Definition A.2.** *A measure $\mu$ on $(\Omega, \mathcal{F})$ is called* **$\sigma$-finite** *if $\Omega$ is the countable union of measurable sets $\Omega_i$'s with $\mu(\Omega_i) < \infty$ for each $i$.*

We proceed to the Radon-Nikodym Theorem.

**Theorem A.3.** ***(Radon-Nikodym Theorem)*** *Let $\mu$ and $\nu$ be two $\sigma$-finite measures on $(\Omega, \mathcal{F})$, with $\nu \ll \mu$. Then there exists a measurable function $h : \Omega \to [0, \infty)$ such that*

$$\nu(A) = \int_A h \, d\mu, \ \forall \, A \in \mathcal{F}.$$

*Moreover the function $h$ is unique in the sense that if there is another measurable function $g : \Omega \to [0, \infty)$ also satisfying this condition, then $h = g$ a.e. ($\mu$).*

The function $h$ is called the Radon-Nikodym derivative of $\nu$ with respect to $\mu$; we write $h = d\nu/d\mu$. This derivative also satisfies an important property stated in Theorem A.4.

**Theorem A.4.** *The function $h$ defined in the Radon-Nikodym theorem is the unique function (up to a.e. ($\mu$)) that satisfies*

$$\int f \, d\nu = \int f h \, d\mu,$$

*for all measurable functions $f$ in which either integral is well defined.*

# Appendix B

## B.1 Proof of Theorems 3.2

Assume $O = \overline{O}_{T+1} \in \mathbb{R}^n$. Consider an arbitrary Borel subset $U$ of $\mathbb{R}^n$. We have

$$\mathbf{1}_{\{O \in U\}} = \sum_{d'} \left( \mathbf{1}_{\{O^{d'} \in U\}} \cdot \prod_{k=0}^{T} \mathbf{1}_{\{w_k = d'_k(X_0)\}} \right),$$

which implies

$$\mathbf{1}_{\{O \in U\}} \cdot s_T^d(X_0) = \sum_{d'} \left( \mathbf{1}_{\{O^{d'} \in U\}} \cdot \prod_{k=0}^{T} \mathbf{1}_{\{w_k = d'_k(X_0)\}} \right) \cdot \prod_{k=0}^{T} \frac{\mathbf{1}_{\{w_k = d_k(X_0)\}}}{e_{d_k(X_0)}(X_0)} = \mathbf{1}_{\{O^d \in U\}} \cdot \prod_{k=0}^{T} \frac{\mathbf{1}_{\{w_k = d_k(X_0)\}}}{e_{d_k(X_0)}(X_0)}.$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[ \mathbf{1}_{\{O \in U\}} \cdot s_T^d(X_0) \middle| X_0 \right] &= \mathbb{E}\left[ \mathbf{1}_{\{O^d \in U\}} \cdot \prod_{k=0}^{T} \frac{\mathbf{1}_{\{w_k = d_k(X_0)\}}}{e_{d_k(X_0)}(X_0)} \middle| X_0 \right] \\
&= \frac{1}{\prod_{k=0}^{T} e_{d_k(X_0)}(X_0)} \cdot \mathbb{E}\left[ \mathbf{1}_{\{O^d \in U\}} \cdot \prod_{k=0}^{T} \mathbf{1}_{\{w_k = d_k(X_0)\}} \middle| X_0 \right] \\
&= \frac{1}{\prod_{k=0}^{T} e_{d_k(X_0)}(X_0)} \cdot \mathbb{E}\left[ \mathbf{1}_{\{O^d \in U\}} \middle| X_0 \right] \cdot \prod_{k=0}^{T} \mathbb{E}\left[ \mathbf{1}_{\{w_k = d_k(X_0)\}} \middle| X_0 \right].
\end{aligned}$$

This implies

$$\mathbb{E}\left[ \mathbf{1}_{\{O \in U\}} \cdot s_T^d(X_0) \middle| X_0 \right] = \mathbb{E}\left[ \mathbf{1}_{\{O^d \in U\}} \middle| X_0 \right].$$

Here, the third equality holds because of Assumption 3.1. Since this result holds for any Borel subset $U$ of $\mathbb{R}^n$, we conclude that $s_T^d(X_0)$ is a Radon-Nikodym derivative of $P_d^m(O^d|X_0)$ with respect to $P^m(O|X_0)$. By Theorem A.4, we are done.

## B.2 Proof of Theorem 3.3

Most of the proof comes from Orellana, Rotnitzky, and Robins [26]. We supply it here with detailed explanations in each step. We start with an important lemma.

**Lemma B.1.** *For each $j \in \{0, 1, ..., T\}$, we have*

$$\mathbb{E}\left[ \underline{m}_{j-1,T}\left( \overline{O}_T^d, \left( \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T} \right) \right) \middle| O^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}) \right] = 1.$$

**Proof:** We first define $\underline{m}_{j,j}\left(\overline{O}_j^d, \left(\overline{d}_j(\overline{O}_j^d, \overline{W}_{j-1}), \underline{W}_{j,j}\right)\right) = 1$ for $j \in \{0, 1, ..., T\}$. Then

$$
\begin{aligned}
& \mathbb{E}\left[\underline{m}_{j-1,T}\left(\overline{O}_T^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,S}\right)\right) \middle| O^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right] \\
= \; & \mathbb{E}\left[\underline{m}_{T-1,T}\left(\overline{O}_T^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,S}\right)\right) \middle| O^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right] \\
& \times \underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,S-1}\right)\right) \qquad \text{(B.1)} \\
= \; & \mathbb{E}\left[\underline{m}_{T-1,T}\left(\overline{O}_T^d, \overline{W}_T\right) \middle| O^d, \overline{W}_{T-1} = \overline{d}_{T-1}(\overline{O}_{T-1}^d, \overline{W}_{T-2})\right] \\
& \times \underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right)\right) \qquad \text{(B.2)} \\
= \; & \mathbb{E}\left[\frac{\mathbf{1}_{\{W_T = d_T(\overline{O}_T^d, \overline{W}_{T-1})\}}}{e_{d_T}(\overline{O}_T^d, \overline{d}_{T-1}(\overline{O}_{T-1}^d, \overline{W}_{T-2}))} \middle| O^d, \overline{W}_{T-1} = \overline{d}_{T-1}(\overline{O}_{S-1}^d, \overline{W}_{T-2})\right] \\
& \times \underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right)\right) \qquad \text{(B.3)} \\
= \; & \frac{\mathbb{E}\left[\mathbf{1}_{\{W_T = d_T(\overline{O}_T^d, \overline{W}_{T-1})\}} \middle| O^d, \overline{W}_{T-1} = \overline{d}_{T-1}(\overline{O}_{T-1}^d, \overline{W}_{T-2})\right]}{e_{d_T}(\overline{O}_T^d, \overline{d}_{T-1}(\overline{O}_{T-1}^d, \overline{W}_{T-2}))} \\
& \times \underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,S-1}\right)\right) \qquad \text{(B.4)} \\
= \; & \frac{\mathbb{P}\left[W_T = d_T(\overline{O}_T^d, \overline{W}_{T-1}) \middle| O^d, \overline{W}_{T-1} = \overline{d}_{T-1}(\overline{O}_{T-1}^d, \overline{W}_{T-2})\right]}{e_{d_T}(\overline{O}_T^d, \overline{d}_{T-1}(\overline{O}_{T-1}^d, \overline{W}_{T-2}))} \\
& \times \underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right)\right) \\
= \; & \frac{\mathbb{P}\left[W_T = d_T(\overline{O}_T, \overline{W}_{T-1}) \middle| O_{T+1}^d, \overline{O}_T, \overline{W}_{T-1} = \overline{d}_{T-1}(\overline{O}_{T-1}, \overline{W}_{T-2})\right]}{e_{d_T}(\overline{O}_T, \overline{d}_{T-1}(\overline{O}_{T-1}, \overline{W}_{T-2}))} \\
& \times \underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right)\right) \qquad \text{(B.5)} \\
= \; & \frac{\mathbb{P}\left[W_T = d_T(\overline{O}_T, \overline{W}_{T-1}) \middle| \overline{O}_T, \overline{W}_{T-1} = \overline{d}_{T-1}(\overline{O}_{T-1}, \overline{W}_{T-2})\right]}{e_{d_T}(\overline{O}_T, \overline{d}_{T-1}(\overline{O}_{T-1}, \overline{W}_{T-2}))} \\
& \times \underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right)\right) \qquad \text{(B.6)} \\
= \; & \underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right)\right).
\end{aligned}
$$

Here, equation (B.1) holds because $\underline{m}_{j-1,T-1}(\cdot,\cdot)$ is a constant conditional on the obervables. Equation (B.2) holds by first directly plugging in the conditional part in the expectation; then given $\overline{O}_{T+1}^d$, the unit must follow regime $d$ so $\overline{W}_{T-1} = \overline{d}_{T-1}(\overline{O}_{T-1}^d, \overline{W}_{S-2})$. Equation (B.3) holds by definition of $\underline{m}_{T-1,T}(\cdot,\cdot)$. Equation (B.4) holds because we assume the value $e_{d_T(\overline{O}_T^d)}(\cdot,\cdot)$ is estimated before fitting in the model so it is a constant. Equation (B.5) holds because by Assumption 2.1, the event $(\overline{O}_T = \overline{o}_T \;\&\; \overline{W}_{T-1} = \overline{d}_{T-1}(\overline{o}_{T-1}, \overline{W}_{T-2}))$

is equivalent to the event $(\overline{O}_T^d = \overline{o}_T \ \& \ \overline{W}_{T-1} = \overline{d}_{T-1}(\overline{o}_{T-1}, \overline{W}_{T-2}))$. Equation (B.6) holds because by Assumption 2.2, $W_T$ is conditionally independent of $O_{T+1}^d$ given $\overline{O}_T$ and $\overline{W}_{T-1} = \overline{d}_{T-1}(\overline{O}_{T-1}, \overline{W}_{T-2})$.

Now by the law of iterated expectation, we get

$$\begin{aligned}
& \mathbb{E}\left[\underline{m}_{j-1,T}\left(\overline{O}_T^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T}\right)\right) \middle| O^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2})\right] \\
= \ & \mathbb{E}\left[\mathbb{E}\left[\underline{m}_{j-1,T}\left(\overline{O}_T^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d), \underline{W}_{j-1,T}\right)\right) \middle| O^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right] \middle| \right. \\
& \left. O^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2})\right] \\
= \ & \mathbb{E}\left[\underline{m}_{j-1,T-1}\left(\overline{O}_{T-1}^d, \left(\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T-1}\right)\right) \middle| \overline{O}_{T+1}^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2})\right].
\end{aligned}$$

When $j = T$, this value is 1. So the lemma holds for $j = T$. Moreover, we can repeat this process in exactly the same way. In other words, we can work backward one step at a time for $j = T-1, T-2, ...$ until $j = 0$ where in each step (that is, for each value of $j$) we can show that the lemma holds. Therefore, Lemma B.1 is proved. □

Now fix $j \in \{0, 1, ..., T\}$. Still assume $O = \overline{O}_{T+1} \in \mathbb{R}^n$. Let $U$ be an arbitrary Borel subset of $\mathbb{R}^n$. We note that by Assumption 2.1, for each $j$ the following two events are equivalent:

$$\left(\overline{O}_j = \overline{o}_j \ \& \ \overline{W}_{j-1} = \overline{d}_{j-1}\left(\overline{o}_{j-1}, \overline{W}_{j-2}\right)\right) \text{ and } \left(\overline{O}_j^d = \overline{o}_j \ \& \ \overline{W}_{j-1} = \overline{d}_{j-1}\left(\overline{o}_{j-1}, \overline{W}_{j-2}\right)\right).$$

This implies

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{\{O \in U\}} \times \underline{m}_{j-1,T}(\overline{O}_T, \overline{W}_T) \middle| \overline{O}_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}, \overline{W}_{j-2})\right] \\
= \ & \mathbb{E}\Big[\mathbf{1}_{\{(\overline{O}_j^d, \underline{O}_{j,T+1}) \in U\}} \\
& \times \ \underline{m}_{j-1,T}((\overline{O}_j^d, \underline{O}_{j,T}), (\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T})) \Big| \overline{O}_j^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2})\Big].
\end{aligned}$$

Moreover, $\underline{m}_{j-1,T}((\overline{O}_j^d, \underline{O}_{j,T}), (\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T}))$ is zero except when $W_j = d_j(\overline{O}_j^d, \overline{W}_{j-1})$, $W_{j+1} = d_{j+1}((\overline{O}_j^d, O_{j+1}), \overline{W}_j)$, ..., $W_T = d_T((\overline{O}_j^d, \underline{O}_{j,T}), \overline{W}_{T-1})$, that is, when $\underline{W}_{j-1,T}$ also follows regime $d$. In this case

$$\begin{aligned}
& \mathbf{1}_{\{(\overline{O}_j^d, \underline{O}_{j,T+1}) \in U\}} \times \underline{m}_{j-1,T}((\overline{O}_j^d, \underline{O}_{j,T}), (\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T})) \\
= \ & \mathbf{1}_{\{O^d \in U\}} \times \underline{m}_{j-1,T}(\overline{O}_T^d, (\overline{d}_{j-1}(\overline{O}_{j-1}^d, \overline{W}_{j-2}), \underline{W}_{j-1,T})).
\end{aligned}$$

Hence,

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{\{O\in U\}} \times \underline{m}_{j-1,T}(\overline{O}_T, \overline{W}_T)\middle| \overline{O}_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}, \overline{W}_{j-2})\right] \\
= & \mathbb{E}\left[\mathbf{1}_{\{O^d\in U\}} \times \underline{m}_{j-1,T}(\overline{O}^d_T, (\overline{d}_{j-1}(\overline{O}^d_{j-1}, \overline{W}_{j-2}), \underline{W}_{j-1,T}))\middle| \overline{O}^d_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}^d_{j-1}, \overline{W}_{j-2})\right] \\
= & \mathbb{E}\Big[\mathbb{E}\left[\underline{m}_{j-1,T}(\overline{O}^d_T, (\overline{d}_{j-1}(\overline{O}^d_{j-1}, \overline{W}_{j-2}), \underline{W}_{j-1,S}))\middle| O^d, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}^d_{j-1}, \overline{W}_{j-2})\right] \\
& \times \ \mathbf{1}_{\{O^d\in U\}}\middle| \overline{O}^d_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}^d_{j-1}, \overline{W}_{j-2})\Big] \\
= & \mathbb{E}\left[\mathbf{1}_{\{O^d\in U\}}\middle| \overline{O}^d_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}^d_{j-1}, \overline{W}_{j-2})\right] \text{ (by Lemma B.1).}
\end{aligned}$$

We obtain

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{\{O\in U\}} \times \underline{m}_{j-1,T}(\overline{O}_T, \overline{W}_T)\middle| \overline{O}_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}_{j-1}, \overline{W}_{j-2})\right] \\
= & \mathbb{E}\left[\mathbf{1}_{\{O^d\in U\}}\middle| \overline{O}^d_j, \overline{W}_{j-1} = \overline{d}_{j-1}(\overline{O}^d_{j-1}, \overline{W}_{j-2})\right].
\end{aligned}$$

Since this holds for any Borel subset $U$ of $\mathbb{R}^n$, we conclude that $\underline{m}_{j-1,T}(\overline{O}_T, \overline{W}_T)$ is a Radon-Nikodym derivative of $P^m_d(O^d|\overline{O}^d_j, \overline{W}_{j-1})$ with respect to $P^m(O|\overline{O}_j, \overline{W}_{j-1})$. Theorem A.4 implies that we are done.

# Appendix C Multilayer Perceptron

The multilayer perceptron model takes an input vector (which is, the initial covariates), then goes through all the hidden layers, and finally generates an output vector (which is, the outcome) which can be of different length.

Specifically, consider layer $(k+1)$ (not the input layer) in the network. The units, or neurons, from layer $k$ whose values are $x^k_1, ..., x^k_{r-1}$, and $x^k_r$ are fed to this layer. Each unit $i$ in layer $(k+1)$ whose value $x^{k+1}_i$ is calculated based on the values $x^k_j$'s and the corresponding weights of connections from the neurons in layer $k$ to it, say $w^k_{i,1}, ..., w^k_{i,r}$ and a bias term $b^k_i$:

$$x^{k+1}_i = \sum_{j=1}^{r} w^k_{i,j} x^k_j + b^k_i.$$

This value of $x^{k+1}_i$ in turn produces output $g(x^{k+1}_i)$ for some pre-defined function $g(\cdot)$. These outputs $g(x^{k+1}_i)$'s will then be fed to the next layer. This process continues until it hits the output layer on which we obtain an output prediction. The function $g(\cdot)$ used for all hidden layers is usually different from the function $f(\cdot)$ which would be used in the output layer. In

practice, $g(\cdot)$ is usually a sigmoid function such as

$$g(x) = \frac{1}{1+e^{-x}} \text{ or } g(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

or a rectified linear unit (ReLU) $g(x) = \max(x, 0)$. Meanwhile, $f(\cdot)$ is usually a linear function $f(x) = x$ for regression model and softmax function, that is a multinomial function, for multiclass probability estimation. See Figure 4 for reference.

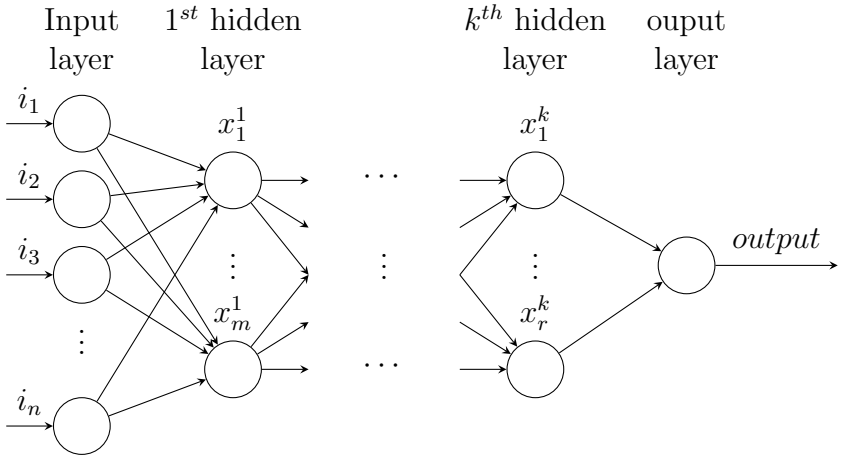



Figure 4: Multilayer Neural Network Model.

Above, we have seen how to get from the input to the estimated output assuming all the parameters (that is, weights and bias terms) are known. What remains is how to estimate them. As in all estimation problems, we would need a training set of data and a loss function for which we choose the parameters to minimize. As mentioned in Section 4, the loss function used here is usually the RMSE.

Since we have a lot of parameters (through many layers) to estimate and the functions $g$ and $f$ used can be complicated, we usually rely on iterative, approximate approaches in contrast to finding the formulas explicitly. The most popular method is *stochastic gradient descent*.[13] The application of the stochastic gradient descent method in this case is quite tricky, however, since we have a huge network structure. Fortunately, computer scientists have worked this out and proposed an algorithm called *backpropagation* (See Rumelhart, Hinton, and Williams [31]). In essence, this algorithm is a computationally efficient way to calculate the gradient of such a large function. Hence, we can effeciently estimate the parameters as we update parameters back and forth when minimizing the loss function. The

[13]This method unfortunately does not guarantee a global minimum for the loss function since it is not necessarily convex.

detailed description of this algorithm is provided in Le [17].

Also mentioned in Section 4, we often use a hyperparameter $\lambda$ to control the parameters; this is a form of regularization. Usually, the function of parameters controlled by $\lambda$ is the sum of squares of all weights $w_{i,j}^k$'s in the network. In practice, people do not regularize the biases since they do not interact with the data through multiplicative interactions and thus should not control the data dimension's impact on the final objective function. We note that the backpropagation algorithm can be modified straightforwardly when we add the regularization term to the loss function.

In general, we follow the procedure described in Section 4 to find the optimal model. We first divide the data set into the training, the validation, and the test sets. For each value of the hyperparameter $\lambda$, we use cross-validation on the training set to choose the optimal number of hidden layers and the optimal number of neurons in each hidden layer. We use the validation set to choose the optimal $\lambda$. Finally, we evaluate the performance of the derived model using the test set.

# Appendix D Statistical Properties and Inference

Consider the static treatment regime setting, specifically Theorem 3.2, while noting that the dynamic case is similar.

Assume that we want to estimate $\tau(X_0) = \mathbb{E}[u(O^d) - u(O^{d'})|X_0]$. Here, $d$ and $d'$ are two pre-assigned treatment regimes. Let $Y^{TO}$ be the transformed outcome. Then

$$\mathbb{E}[Y^{TO}|X_0] = \tau(X_0).$$

Assume using Multilayer Perceptron gives us the estimator $\widehat{\tau}(X_0)$. We have, conditional on $X_0$, that

$$\begin{aligned}\mathbb{E}\left[(Y^{TO} - \widehat{\tau}(X_0))^2\right] = \mathbb{E}\left[(Y^{TO} - \tau(X_0))^2\right] + \mathbb{E}\left[(\tau(X_0) - \widehat{\tau}(X_0))^2\right] \\ + 2 \times \mathbb{E}\left[(Y^{TO} - \tau(X_0))(\tau(X_0) - \widehat{\tau}(X_0))\right]\end{aligned}$$

where the last term is

$$2\left(\tau(X_0) - \widehat{\tau}(X_0)\right) \cdot \mathbb{E}\left[Y^{TO} - \tau(X_0)|X_0\right] = 2\left(\tau(X_0) - \widehat{\tau}(X_0)\right) \cdot \left(\mathbb{E}\left[Y^{TO}|X_0\right] - \tau(X_0)\right) = 0.$$

Hence

$$\mathbb{E}\left[(Y^{TO} - \widehat{\tau}(X_0))^2\right] = \mathbb{E}\left[(Y^{TO} - \tau(X_0))^2\right] + \mathbb{E}\left[(\tau(X_0) - \widehat{\tau}(X_0))^2\right].$$

If we did not use regularization, then the Multilayer Perceptron technique that we use would have produced $\widehat{\tau}(\cdot)$ that minimizes the empirical approximation of $\mathbb{E}[(Y^{TO} - \widehat{\tau}(X_0))^2]$. Because $\widehat{\tau}(\cdot)$ is not contained in $\mathbb{E}[(Y^{TO} - \tau(X_0))^2]$, it would be chosen to minimize $\mathbb{E}[(\tau(X_0) - \widehat{\tau}(X_0))^2]$. This term is minimized when $\widehat{\tau}(\cdot) \equiv \tau(\cdot)$. In other words, we would obtain an unbiased estimator for $\tau(\cdot)$. This result is formally stated in Theorem D.1.

**Theorem D.1.** *With no use of regularization, $\widehat{\tau}$ is an unbiased estimator for $\tau$.*

There is a trade-off between bias and variance. In our framework, we use regularization to reduce variance but at the same time make $\widehat{\tau}$ biased. Also, $\widehat{\tau}$ would not necessarily be consistent. There are some results regarding the consistency and asymptotic normality of estimators in linear models when regularization is present (See Knight and Fu [16]). However, such results in deep learning or machine learning in general are still absent.

Now to draw statistical inference and do hypothesis testing, for a new realized covariate vector $x_0^{new}$ we need to (consistently) estimate $Var(u(O^d) - u(O^{d'})|x_0^{new})$.

To this end, we recall that Newey [24] proposes the kernel approach to estimate the variance. Specifically, we consider the model $Y = g(X) + \epsilon$ with observed outcome $Y$ and $\mathbb{E}[\epsilon|X] = 0$. Let $\widehat{g}(\cdot)$ be an estimator for $g(\cdot)$ and $(X_i, Y_i)_{i=1}^n$ be $n$ observations. Then a non-parametric estimator for the variance $\sigma^2(x) = Var(Y|X = x)$ is

$$\widehat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} K(H^{-1}(X_i - x))\widehat{\epsilon}_i^2}{\sum_{i=1}^{n} K(H^{-1}(X_i - x))},$$

where $\widehat{\epsilon}_i = Y_i - \widehat{g}(X_i)$ for each $i$, $K(\cdot)$ is a multivariate kernel function, and $H$ is the bandwidth matrix.

The intuition why this approach is asymptotically valid is that the parametric calculation accounts correctly for the variance of the estimator and that the bias shrinks faster than the standard deviation in the asymptotic approximation. Fan and Yao [8] verify this intuition by showing the asymptotic normality property of $\widehat{\sigma}^2(x)$ as an estimator for $\sigma^2(x)$ for each $x$.

Applying the same method to our setting is not so straightforward as the ground truth is not known. Similarly to how we deal with the testing methodology, we propose a matching based kernel approach to study the statistical inference. We also define $M$ units in the test set with covariates $x_0^{d,m}$ and outcome $u(o^{d,m})$ and $M$ matching units also in the test set with covariates $x_0^{d',m}$ and outcome $u(o^{d',m})$ for $m = 1, ..., M$. Assume $x_0^{new}$ is in the test set. To estimate $\sigma^2(x_0^{new}) = Var(u(O^d) - u(O^{d'})|x_0^{new})$, we first define for each $m \in \{1, ..., M\}$

$$\widehat{\epsilon}_m = u(o^{d,m}) - u(o^{d',m}) - \frac{1}{2}\left(\widehat{\tau}(x_0^{d,m}) + \widehat{\tau}(x_0^{d',m})\right).$$

This is exactly $(\widetilde{\tau}_m - \widehat{\tau}_m)$ in Section 5. Then, an estimate for $\sigma^2(x_0^{new})$ is

$$\widehat{\sigma}^2(x_0^{new}) = \frac{\sum_{m=1}^{M} K\left(H^{-1}\left[\frac{x_0^{d,m}+x_0^{d',m}}{2} - x_0^{new}\right]\right)\widehat{\epsilon}_m^2}{\sum_{m=1}^{M} K\left(H^{-1}\left[\frac{x_0^{d,m}+x_0^{d',m}}{2} - x_0^{new}\right]\right)}. \tag{D.1}$$

When the data set is large enough, we can choose almost identical pairs $(x_0^{d,m}, x_0^{d',m})$. Thus, we obtain the "observed" outcome $u(o^{d,m}) - u(o^{d',m})$ for each $x_0^{d,m}$. However, $\widehat{\tau}(\cdot)$ is an estimator for $\tau(\cdot)$, which is obtained by using transformed outcome $Y^{TO}$ instead of the real outcome $u(O^d) - u(O^{d'})$. So even though on average these two outcomes are equal, the real bias induced by this discrepancy can be large. As a result, we cannot guarantee that the bias shrinks faster than the standard deviation in the asymptotic approximation. In other words, we cannot guarantee that $\widehat{\sigma}^2$ is asymptotically normal.

However, it is an indicative and reasonable estimator. To see why, we first write for each $m = 1, ..., M$ that

$$W_m(x_0^{new}) = K\left(H^{-1}\left[\frac{x_0^{d,m}+x_0^{d',m}}{2} - x_0^{new}\right]\right) \Big/ \sum_{i=1}^{M} K\left(H^{-1}\left[\frac{x_0^{d,i}+x_0^{d',i}}{2} - x_0^{new}\right]\right);$$

$$\widehat{\tau}_m = \frac{1}{2}\left(\widehat{\tau}(x_0^{d,m}) + \widehat{\tau}(x_0^{d',m})\right); y_m^{True} = \left(u(o^d) - u(o^{d'})\right)_m;$$

and $y_m^{TO}$ as the $m^{th}$ transformed outcome value. Then

$$\begin{aligned}
\widehat{\sigma}^2(x_0^{new}) &= \sum_{m=1}^{M} W_m(x_0^{new})\left(y_m^{True} - \widehat{\tau}_m\right)^2 \\
&= \sum_{m=1}^{M} W_m(x_0^{new})\left[(y_m^{True} - y_m^{TO})^2 + (y_m^{TO} - \widehat{\tau}_m)^2 + 2(y_m^{True} - y_m^{TO})(y_m^{TO} - \widehat{\tau}_m)\right] \\
&= \sum_{m=1}^{M} W_m(x_0^{new})(y_m^{True} - y_m^{TO})^2 + \sum_{m=1}^{M} W_m(x_0^{new})(y_m^{TO} - \widehat{\tau}_m)^2 + \\
&\qquad 2\sum_{m=1}^{M} W_m(x_0^{new})(y_m^{True} - y_m^{TO})y_m^{TO} - 2\sum_{m=1}^{M} W_m(x_0^{new})(y_m^{True} - y_m^{TO})\widehat{\tau}_m.
\end{aligned}$$

If the following results hold[14]

[14]They hold with the standard kernel method setting. In this setting, in order for them to hold we need to make additional assumptions due to the discrepancy in matching.

$$\sum_{m=1}^{M} W_m(x_0^{new})(y_m^{True} - y_m^{TO})^2 \xrightarrow{p} \mathbb{E}\left[(Y^{True} - Y^{TO})^2 | x_0^{new}\right];$$

$$\sum_{m=1}^{M} W_m(x_0^{new})(y_m^{TO} - \widehat{\tau}_m)^2 \xrightarrow{p} Var(Y^{TO}|x_0^{new});$$

$$\sum_{m=1}^{M} W_m(x_0^{new})(y_m^{True} - y_m^{TO})y_m^{TO} \xrightarrow{p} \mathbb{E}\left[(Y^{True} - Y^{TO})Y^{TO}|x_0^{new}\right];$$

$$\sum_{m=1}^{M} W_m(x_0^{new})(y_m^{True} - y_m^{TO})\widehat{\tau}_m \xrightarrow{p} \mathbb{E}\left[(Y^{True} - Y^{TO})\widehat{\tau}|x_0^{new}\right] \to 0;$$

then

$$\widehat{\sigma}^2(x_0^{new}) \xrightarrow{p} \mathbb{E}\left[(Y^{True} - Y^{TO})^2|x_0^{new}\right] + Var(Y^{TO}|x_0^{new}) + 2\mathbb{E}\left[(Y^{True} - Y^{TO})Y^{TO}|x_0^{new}\right].$$

On the other hand,

$$\begin{aligned}
\sigma^2(x_0^{new}) &= Var(Y^{True}|x_0^{new}) \\
&= Var(Y^{True} - Y^{TO}|x_0^{new}) + Var(Y^{TO}|x_0^{new}) + 2Cov(Y^{True} - Y^{TO}, Y^{TO}|x_0^{new}) \\
&= \mathbb{E}\left[(Y^{True} - Y^{TO})^2|x_0^{new}\right] + Var(Y^{TO}|x_0^{new}) + 2\mathbb{E}\left[(Y^{True} - Y^{TO})Y^{TO}|x_0^{new}\right].
\end{aligned}$$

This means we can then show that $\widehat{\sigma}^2(x_0^{new})$ defined in D.1 is a consistent estimator for $\sigma^2(x_0^{new}) = Var(u(O^d) - u(O^{d'})|x_0^{new})$.

In practice, $K(\cdot)$ is usually the Multivariate Gaussian density function. To choose the optimal bandwidth matrix $H$, Fan and Yao [8] suggest using cross validation. We can follow them, or simply use Silverman's rule-of-thumb (Silverman [33]) that $H_{ij} = 0$ for $i \neq j$ and

$$H_{ii} = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} M^{\frac{-1}{d+4}} \sigma_i,$$

where $\sigma_i$ is the standard deviation of the $i^{th}$ variable and $d$ is the dimension of $X_0$.

Note however that the same formula may not be valid for units in the validation or training sets because these data are used to estimate the parameters in the model.

For the dynamic treatment regime, we can apply the above procedure in each period, though the statistical inference in period $T = 0$ might be of the most interest.