Journal of Machine Learning Research 24 (2023) 1-92 Submitted 7/21; Revised 10/22; Published 2/13

Jump Interval-Learning for Individualized Decision Making with Continuous Treatments

Hengrui Cai[*] HENGRC1@UCI.EDU
Department of Statistics
University of California Irvine
Irvine, CA 92697, USA

Chengchun Shi[*] C.SHI7@LSE.AC.UK
Department of Statistics
London School of Economics and Political Science
London, WC2A 2AE, UK

Rui Song RSONG@NCSU.EDU
Wenbin Lu WLU4@NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC 27695, USA

Editor: Ji Zhu

Abstract

An individualized decision rule (IDR) is a decision function that assigns each individual a given treatment based on his/her observed characteristics. Most of the existing works in the literature consider settings with binary or finitely many treatment options. In this paper, we focus on the continuous treatment setting and propose a jump interval-learning to develop an individualized interval-valued decision rule (I2DR) that maximizes the expected outcome. Unlike IDRs that recommend a single treatment, the proposed I2DR yields an interval of treatment options for each individual, making it more flexible to implement in practice. To derive an optimal I2DR, our jump interval-learning method estimates the conditional mean of the outcome given the treatment and the covariates via jump penalized regression, and derives the corresponding optimal I2DR based on the estimated outcome regression function. The regressor is allowed to be either linear for clear interpretation or deep neural network to model complex treatment-covariates interactions. To implement jump interval-learning, we develop a searching algorithm based on dynamic programming that efficiently computes the outcome regression function. Statistical properties of the resulting I2DR are established when the outcome regression function is either a piecewise or continuous function over the treatment space. We further develop a procedure to infer the mean outcome under the (estimated) optimal policy. Extensive simulations and a real data application to a Warfarin study are conducted to demonstrate the empirical validity of the proposed I2DR.

Keywords: Continuous treatment, Dynamic programming, Individualized interval-valued decision rule, Jump interval-learning, Precision medicine

[*]. Equal contribution.

1. Introduction

Individualized decision making is an increasingly attractive artificial intelligence paradigm that proposes to assign each individual a given treatment based on their observed characteristics. In particular, such a paradigm has been recently employed in precision medicine to tailor the individualized treatment decision rule. Among all individualized decision rules (IDR), the one that maximizes the expected outcome is referred to as an optimal IDR. There is a huge literature on learning the optimal decision rule. Some popular methods include Q-learning (Watkins and Dayan, 1992; Chakraborty et al., 2010; Qian and Murphy, 2011; Song et al., 2015), A-learning (Robins, 2004; Murphy, 2003; Shi et al., 2018), policy search methods (Zhang et al., 2012, 2013; Wang et al., 2018; Nie et al., 2020), outcome-weighted learning (Zhao et al., 2012, 2015; Zhu et al., 2017; Meng et al., 2020), concordance-assisted learning (Fan et al., 2017; Liang et al., 2017), decision list-based methods (Zhang et al., 2015, 2018), and direct learning (Qi et al., 2020). We note, however, all these methods consider settings where the number of available treatment options is finite.

In this paper, we consider individualized decision making in continuous treatment settings. These studies occur in a number of real applications, including personalized dose finding (Chen et al., 2016) and dynamic pricing (den Boer and Keskin, 2020). For instance, in personalized dose finding, one wishes to derive a dose level or dose range for each patient. Due to patients' heterogeneity in response to doses, it is commonly assumed that there may not exist a unified best dose for all patients. Thus, one major interest in precision medicine is to develop an IDR that assigns each individual patient a certain dose level or a specified range of doses based on their individual personal information, to optimize their health status. Similarly, in dynamic pricing, we aim to identify an IDR that assigns each product an optimal price according to its characteristics to maximize the overall profit.

In contrast to developing the optimal IDR under discrete treatment settings, individualized decision making with a continuous treatment domain has been less studied. Among those available, Rich et al. (2014) modeled the interactions between the dose level and covariates to recommend personalized dosing strategies. Laber and Zhao (2015) developed a tree-based method to derive the IDR by dividing patients into subgroups and assigning each subgroup the same dose level. Chen et al. (2016) proposed an outcome-weighted learning method to directly search the optimal IDR among a restricted class of IDRs. Kallus and Zhou (2018) and Chernozhukov et al. (2019) evaluated and optimized IDRs for continuous treatments by replacing the indicator function in the doubly-robust approach with the kernel function, and by modeling the conditional mean outcome function (i.e., the value) through a semi-parametric form, respectively. Zhu et al. (2020) focused on the class of linear IDRs and proposed to compute an optimal linear IDR by maximizing a kernel-based value estimate. Schulz and Moodie (2020) proposed a doubly robust estimation method for personalized dose finding. Zhou et al. (2021) proposed a dimension reduction framework for personalized dose finding that effectively reduces the dimensionality of baseline characteristics from a high to a moderate scale. The estimated optimal IDRs computed by these methods typically recommend one single treatment level for each individual, making it hard to implement in practice.

The focus of this paper is to develop an individualized interval-valued decision rule (I2DR) that returns a range of treatment levels based on individuals' baseline information.

Compared to the IDRs recommended by the existing works, the proposed I2DR is more flexible to implement in practice. Take personalized dose finding as an illustration. First, interval-valued dose levels may be applied to patients of the same characteristics, when an arbitrary dose within the given dose interval could achieve the same efficacy. Studies of the pharmacokinetics of vancomycin conducted by Rotschafer et al. (1982) suggested that adults with normal renal function should receive an initial dosage of 6.5 to 8 milligrams of vancomycin per kilogram intravenously over 1 hour every 6 to 12 hours. In the review of Warfarin dosing reported by Kuruvilla and Gurk-Turner (2001), when the international normalized ratio (INR) approaches the target range or omits dose, they suggested giving 1-2.5 milligram vitamin K1 if a patient has a risk factor for bleeding, otherwise provide Vitamin K1 2-4 milligram orally. Second, in cases where the available dose levels are limited, recommending a single dose is not practical. The proposed interval-valued dose rule gives more options. Based on the proposed interval, the decision maker can select the most appropriate dose by taking some other factors (e.g., patient affordability or side effects) into consideration. Third, a range of doses gives instructions for designing the medicine specification and helps to save costs on manufacturing dosage. Finally, many medical applications including treating chronic disease (Flack and Adekola, 2020) and radiation therapy for cancer (Scott et al., 2017) prefer optimal dose interval recommendation.

Our contributions are summarized as follows. Scientifically, individualized decision making in a continuous treatment domain is a vital problem in many applications such as precision medicine and dynamic pricing. To the best of our knowledge, this is the first work on developing individualized interval-valued decision rules. Our proposal thus fills a crucial gap, extends the scope of existing methods that focus on recommending IDRs, and offers a useful tool for individualized decision making in a number of applications.

Methodologically, we propose a novel jump interval-learning (JIL) by integrating personalized decision making with multi-scale change point detection (see Niu et al., 2016, for a selective overview). Our proposal makes useful contributions to the two aforementioned areas simultaneously.

First, to implement personalized decision making, we propose a data-driven I2DR in a continuous treatment domain. Our proposal is motivated by the empirical finding that the expected outcome can be a piecewise function in the treatment domain in various applications. Specifically, in dynamic pricing (den Boer and Keskin, 2020), the expected demand (outcome Y of interest) for a product has jump discontinuities as a function of the charged price (action A) and baseline information such as income (covariates X). In other words, a small price change will lead to a considerably different demand given fixed covariates. In these applications, it is reasonable to impose a piecewise-function model for the outcome regression function. We then leverage ideas from the change point detection literature and propose a jump-penalized regression to estimate the conditional mean of the expected outcome as a function of the treatment level and the baseline characteristics (outcome regression function). This partitions the entire treatment space into several subintervals. The proposed I2DR is a set of decision rules that assign each subject to one of these subintervals. In addition, we further develop a procedure to construct a confidence interval (CI) for the expected outcome under the proposed I2DR and the optimal IDR.

Second, we note that most works in the multi-scale change point detection literature either focused on models without covariates, or required the underlying truth to be piecewise constant (see e.g., Boysen et al., 2009; Frick et al., 2014; Fryzlewicz, 2014, and the references therein). Our work goes beyond those cited above in that we consider a more complicated (nonparametric) model with covariates, and allow the underlying outcome regression function to be either a piecewise or continuous function over the treatment space. To approximate the expected outcome as a function of baseline covariates, we propose a linear function model and a deep neural networks model. We refer to the two procedures as L-JIL and D-JIL, respectively. Here, the proposed L-JIL yields a set of linear decision rules that is easy to interpret. See the real data analysis in Section 6 for details. On the contrary, the proposed D-JIL employs deep learning (LeCun et al., 2015) to model the complicated outcome-covariates relationships that often occur in high-dimensional settings. We remark that both procedures are developed by imposing a piecewise-function model to approximate the outcome-treatment relationship. Yet, they are valid when the expected outcome is a continuous function of the treatment level as well.

Theoretically, we systematically study the statistical properties of the jump-penalized estimators with linear regression or deep neural networks. Our theoretical approaches can be applied to the analysis of general covariate-based change point models. The model could be either parametric or nonparametric. Specifically, we establish the almost sure convergence rates of our estimators. When the underlying outcome regression function is a piecewise function of the treatment, we further derive the almost sure convergence rates of the estimated change point locations, and show that with probability 1, the number of change points can be correctly estimated with sufficiently large sample size. These findings are nontrivial extensions of classical results derived for models without covariates. For instance, deriving the asymptotic behavior of change point estimators for these models typically relies on the tail inequalities for the partial sum process (see e.g., Frick et al., 2014). However, these technical tools are not directly applicable to our settings where deep learning is adopted to model the outcome regression function. Moreover, we expect our theories to also be of general interest to the line of work on developing theories for deep learning methods (see e.g., Imaizumi and Fukumizu, 2019; Schmidt-Hieber et al., 2020; Farrell et al., 2021).

The rest of this paper is organized as follows. In Section 2, we introduce the statistical framework, define the notion of I2DR, and posit our working model assumptions. In Section 3, we propose the jump interval-learning method and discuss its detailed implementation. Statistical properties of the proposed I2DR and the estimator for the mean outcome under the proposed I2DR are presented in Section 4. We further develop a confidence interval for the expected outcome under the estimated I2DR. Simulation studies are conducted in Section 5 to evaluate the finite sample performance of our proposed method. We apply our method to a real dataset from a Warfarin study in Section 6, followed by a concluding discussion in Section 7. All the proofs are provided in the supplementary article. An R package implementing our proposed I2DR is available on CRAN at https://cran.r-project.org/web/packages/JQL/index.html.

2. Statistical Framework

This section is organized as follows. We first introduce the model setup in Section 2.1. The definition of I2DR is formally presented in Section 2.2. In Section 2.3, we posit two working model assumptions for the expected outcome as a function of the treatment level. We aim to develop a method that works under both working assumptions.

2.1 Model Setup

We begin with some notations. Let A denote the treatment level assigned to a randomly selected individual in the population from a compact interval. Without loss of generality, suppose A belongs to $[0,1]$. Let $X \in \mathbb{X}$ be that individual's baseline covariates where the support $\mathbb{X}$ is a subset in $\mathbb{R}^p$. We assume the covariance matrix of X is positive definite. Let $Y \in \mathbb{R}$ denote that individual's associated outcome, the larger the better by convention. Let $p(\bullet|x)$ denote the probability density function of A given $X = x$. In addition, for any $a \in [0,1]$, define the potential outcome $Y^*(a)$ as the outcome of that individual that would have been observed if they were receiving treatment a. The observed data consists of the covariate-treatment-outcome triplets $\{(X_i, A_i, Y_i) : i = 1, \ldots, n\}$ where (X_i, A_i, Y_i)'s are i.i.d. copies of (X, A, Y). Based on this data, we wish to learn an optimal decision rule to possibly maximize the expected outcome of future subjects using their baseline information.

Formally speaking, an individualized decision rule (IDR) is a deterministic function $d(\cdot)$ that maps the covariate space $\mathbb{X}$ to the treatment space $[0,1]$. The optimal IDR is defined to maximize the expected outcome (value function) $V(d) = \mathrm{E}\{Y^*(d(X))\}$ among all IDRs. The following assumptions guarantee the optimal IDR is identifiable from the observed data.

(A1.) Consistency: $Y = Y^*(A)$, almost surely,
(A2.) No unmeasured confounders: $\{Y^*(a) : a \in [0,1]\} \perp\!\!\!\perp A \mid X$,
(A3.) Positivity: there exists some constant $c_* > 0$ such that $p(a|x) \geq c_*$ for any $x \in \mathbb{X}$ and $a \in [0,1]$.

Assumption (A1) requires the observed outcome to be the same as the potential outcome associated with the observed treatment. Assumption (A2) requires the baseline covariates to contain enough confounders given that the treatment is conditionally independent of the potential outcomes. Assumption (A3) requires the propensity score to be strictly positive for any realization of the baseline covariates. Assumptions (A2) and (A3) automatically hold in randomized studies. In observational studies, one can estimate the propensity score from the data to check (A3). Nonetheless, (A2) cannot be verified in general. In addition, Assumptions (A1) to (A3) are commonly imposed in the literature (see e.g., Chen et al., 2016; Zhu et al., 2020; Schulz and Moodie, 2020) to guarantee that the outcome of interest and the optimal IDR are estimable from observed data. In particular, under (A1)-(A3), we have $V(d) = \mathrm{E}\{Q(X, d(X))\}$ where $Q(x, a) = \mathrm{E}(Y|X = x, A = a)$ is the conditional mean of an individual's outcome given their received treatment and baseline covariates. We refer to this function as the outcome regression function. As a result, the optimal IDR for an individual with covariates x is given by $\arg\max_{a\in[0,1]} Q(x, a)$. Let V^{opt} denote the value function under the optimal IDR. We have $V^{opt} = \mathrm{E}\{\sup_{a\in[0,1]} Q(X, a)\}$.

2.2 I2DR

The focus of this paper is to develop an optimal individualized interval-based decision rule (I2DR). As commented in the introduction, these decision rules are more flexible to implement in practice when compared to single-valued decision rules in personalized dose finding and dynamic pricing.

We define an I2DR as a function $d(\cdot)$ that takes an individual's covariates x as input and outputs an interval $\mathcal{I} \subseteq [0,1]$. Given the recommended interval $\mathcal{I}$, different doctors/agents might assign different treatments to patients/products according to their own preferences. In practice, the decision maker could take the minimum value, the maximum value, the mid-point value, or the value uniformly at random. The actual treatments that subjects receive in the population will have a distribution function $\Pi^*(\cdot; x, \mathcal{I})$. Throughout this paper, we assume $\Pi^*(\cdot; x, \mathcal{I})$ has a bounded density function $\pi^*(\cdot; x, \mathcal{I})$ for any x and $\mathcal{I}$. Apparently, we have $\int_{\mathcal{I}} \pi^*(a; x, \mathcal{I}) da = 1$, for any interval $\mathcal{I}$ and $x \in \mathbb{X}$. When (A1)-(A3) hold, the associated value function under an I2DR $d(\cdot)$ equals

$$V^{\pi^*}(d) = \mathrm{E}\left(\int_{d(X)} Q(X, a) \pi^*(a; X, d(X)) da \right).$$

Restricting $d(\cdot)$ to be a scalar-valued function, $V^{\pi^*}(d)$ is reduced to $V(d)$.

Given the dataset, one may estimate $V^{\pi^*}(d)$ nonparametrically for any $d(\cdot)$ and directly search the optimal I2DR based on the estimated value function. However, such a value search method has the following two limitations. First, a nonparametric estimator of $V^{\pi^*}(d)$ requires specifying the preference function π^*, which might be unknown to us. Second, even though a nonparametric estimator of $V^{\pi^*}(d)$ can be derived, it remains unknown how to efficiently compute the I2DR that maximizes the estimated value (see Section 7.2.2 for details). To overcome these limitations, we propose a semiparametric model for the outcome regression function and use a model-assisted approach to derive the optimal I2DR. We formally introduce our method in Section 3.

2.3 Working Model Assumptions

In this section, we introduce two working models for the outcome regression function, corresponding to a piecewise function and a continuous function of the treatment level.

Model I (Piecewise Functions). Suppose

$$Q(x, a) = \sum_{\mathcal{I} \in \mathcal{P}_0} q_{\mathcal{I},0}(x) \mathbb{I}(a \in \mathcal{I}) \qquad \forall x \in \mathbb{X}, a \in [0,1], \tag{1}$$

for some partition $\mathcal{P}_0$ of $[0,1]$ and a collection of continuous functions $(q_{\mathcal{I},0})_{\mathcal{I} \in \mathcal{P}_0}$, where the number of intervals in $\mathcal{P}_0$ is finite. Specifically, a partition $\mathcal{P}$ of $[0,1]$ is defined as a collection of mutually disjoint intervals $\{[\tau_0, \tau_1), [\tau_1, \tau_2), \ldots, [\tau_{K-1}, \tau_K]\}$ for some $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{K-1} < \tau_K = 1$ and some integer $K \geq 1$. Here, we only require any two consecutive q-functions to be different. We do not impose any additional constraints. As commented in our introduction, we expect the above model assumption holds in real-world examples such as dynamic pricing.

Model II (Continuous Functions). Suppose $Q(x, a)$ is a continuous function of a and x, for any $x \in \mathbb{X}$ and $a \in [0,1]$.

We aim to propose a new method that works when either Model I (piecewise function) or Model II (continuous function) holds.

3. Methods

In this section, we first present the proposed jump interval-learning and its motivation in Section 3.1. We next introduce two concrete proposals, i.e., linear jump interval-learning and deep jump interval-learning, to detail our methods in Section 3.2. We then present the dynamic programming algorithm to implement jump interval-learning (see Algorithm 1 for an overview) in Section 3.3. Finally, we provide more details on tuning parameter selection in Section 3.4.

3.1 Jump Interval-learning

We use Model I to present the motivation for our jump interval-learning. In view of (1), any treatment level within an interval $\mathcal{I} \in \mathcal{P}_0$ will yield the same efficacy to a given individual. The optimal I2DR is then given by

$$d^{opt}(x) = \arg\max_{\mathcal{I} \in \mathcal{P}_0} q_{\mathcal{I},0}(x),$$

independent of the preference function π^*. To see this, notice that

$$V^{\pi^*}(d^{opt}) = \mathrm{E}\left(\int_{d^{opt}(X)} \sum_{\mathcal{I} \in \mathcal{P}_0} q_{\mathcal{I},0}(X) \mathbb{I}(a \in \mathcal{I}) \pi^*(a; X, d^{opt}(X)) da \right)$$

$$= \mathrm{E} \sum_{\mathcal{I} \in \mathcal{P}_0} q_{\mathcal{I},0}(X) \mathbb{I}(d^{opt}(X) \in \mathcal{I}) \int_{d^{opt}(X)} \pi^*(a; X, d^{opt}(X)) da.$$

For any I2DR $d(\cdot)$, we have $\int_{d^{opt}(X)} \pi^*(a; X, d^{opt}(X)) da = \int_{d(X)} \pi^*(a; X, d(X)) da = 1$ by definition. It follows that

$$\begin{aligned} V^{\pi^*}(d^{opt}) &= \mathrm{E} \sum_{\mathcal{I} \in \mathcal{P}_0} q_{\mathcal{I},0}(X) \mathbb{I}(d^{opt}(X) \in \mathcal{I}) \int_{d(X)} \pi^*(a; X, d(X)) da \\ &\geq \mathrm{E} \int_{d(X)} \sum_{\mathcal{I} \in \mathcal{P}_0} q_{\mathcal{I},0}(X) \mathbb{I}(a \in \mathcal{I}) \pi^*(a; X, d(X)) da = V^{\pi^*}(d), \end{aligned}$$

where the inequality is due to that $Q(X, d^{opt}(X)) = \sum_{\mathcal{I} \in \mathcal{P}_0} q_{\mathcal{I},0}(X) \mathbb{I}(d^{opt}(X) \in \mathcal{I}) \geq \sum_{\mathcal{I} \in \mathcal{P}_0} q_{\mathcal{I},0}(X) \mathbb{I}(a \in \mathcal{I}) = Q(X, a)$, almost surely for any $a \in [0, 1]$. Therefore, to derive the optimal I2DR, it suffices to estimate $q_{\mathcal{I},0}(\cdot)$. For notation simplicity, in the rest of this paper, we denote $V^{\pi^*}(d)$ by $V(d)$ for any decision rule d.

From now on, we focus on a subset of intervals in $[0, 1]$. By *interval* we always refer to those of the form $[a, b)$ for some $0 \leq a < b < 1$ or $[a, 1]$ for some $0 \leq a < 1$. For any partition $\mathcal{P} = \{[0, \tau_1), [\tau_1, \tau_2), \dots, [\tau_{K-1}, 1]\}$, we use $J(\mathcal{P})$ to denote the set of change point locations, i.e, $\{\tau_1, \tau_2, \dots, \tau_{K-1}\}$, and $|\mathcal{P}|$ to denote the number of intervals in $\mathcal{P}$. Our proposed method yields a partition $\widehat{\mathcal{P}}$ and an I2DR $\widehat{d}(\cdot)$ such that $\widehat{d}(x) \in \widehat{\mathcal{P}}$, $\forall x \in \mathbb{X}$. The number of intervals in $\widehat{\mathcal{P}}$ (denoted by $|\widehat{\mathcal{P}}|$) involves a trade-off. If $|\widehat{\mathcal{P}}|$ is too large, then $\widehat{\mathcal{P}}$ will contain many short intervals, making the resulting decision rule hard to implement in practice. Yet, a smaller value of $|\widehat{\mathcal{P}}|$ might result in a smaller value function. Our proposed method adaptively determines $|\widehat{\mathcal{P}}|$ based on jump-penalized regression.

We next detail our method. Jump interval-learning consists of the following two steps. In the first step, we estimate the outcome regression function using jump penalized least squares regression. Then we derive the corresponding I2DR from the resulting estimator $\widehat{q}_{\mathcal{I}}(\cdot)$. To begin with, we cut the entire treatment range into m initial intervals:

$$[0,1/m),[1/m,2/m),\dots,[(m-1)/m,1]. \tag{2}$$

The integer m is allowed to diverge with the number of observations n. For instance, it can be specified by the clinical physician such that the output dose interval for each individual is at least of the length m^{-1}. When no prior knowledge is available, we recommend setting m to be proportional to n. It is worth mentioning that (2) is not the final partition that we recommend. Nor is it equal to $\mathcal{P}_0$ defined in Model I. Given (2), we are looking for a partition $\widehat{\mathcal{P}}$ such that each interval in $\widehat{\mathcal{P}}$ corresponds to a union of some of the these m intervals. In other words, we will adaptively combine some of these intervals to form $\widehat{\mathcal{P}}$.

More specifically, let $\mathcal{B}(m)$ denote the set of partitions $\mathcal{P}$ that satisfy the following requirement: the end-points of each interval $\mathcal{I}\in\mathcal{P}$ lie on the grid $\{j/m : j=0,1,\dots,m\}$. We associate to each partition $\mathcal{P}\in\mathcal{B}(m)$ a collection of functions $\{q(\cdot;\theta_{\mathcal{I}})\}_{\mathcal{I}\in\mathcal{P}}\in\prod_{\mathcal{I}\in\mathcal{P}}\mathcal{Q}_{\mathcal{I}}$ for $\mathcal{Q}_{\mathcal{I}}$ as some class of functions, where $\theta_{\mathcal{I}}$ is the underlying parameter associated to interval $\mathcal{I}$. We propose to estimate $\widehat{\mathcal{P}}$ by solving

$$(\widehat{\mathcal{P}},\{\widehat{q}_{\mathcal{I}}:\mathcal{I}\in\widehat{\mathcal{P}}\}) = \\ \mathop{\arg\min}_{\substack{\mathcal{P}\in\mathcal{B}(m)\\ \{q(\cdot;\theta_{\mathcal{I}})\in\mathcal{Q}_{\mathcal{I}}:\mathcal{I}\in\mathcal{P}\}}} \left\{\sum_{\mathcal{I}\in\mathcal{P}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\big\{Y_i-q(X_i;\theta_{\mathcal{I}})\big\}^2+\lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}\|_2^2\right)+\gamma_n|\mathcal{P}|\right\}, \tag{3}$$

where λ_n and γ_n are some nonnegative regularization parameters specified in Section 3.4, and $\|\theta_{\mathcal{I}}\|_2^2$ denote the Euclidean norm of the model parameter $\theta_{\mathcal{I}}$. The purpose of introducing the ℓ_2-type penalty term $\lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}\|_2^2$ is to help to prevent overfitting in large p problems. The purpose of introducing the ℓ_0-type penalty term $\gamma_n|\mathcal{P}|$ is to control the total number of jumps. When $m=n$, $\lambda_n=0$, $A_i=i/n,\forall 1\le i\le n$, no baseline covariates are collected, the above optimization corresponds to the jump-penalized least square estimator proposed by Boysen et al. (2009). We refer to this step as jump interval-learning (JIL).

For a fixed $\mathcal{P}$, solving the optimization function in (3) yields its associated outcome regression functions $\{\widehat{q}_{\mathcal{I}}\}_{\mathcal{I}\in\mathcal{P}}$. This step involves parametric or nonparametric regression and can be solved via existing statistical or machine learning approaches. We provide two concrete study cases below, based on linear regression and deep learning. These estimated outcome regression functions can be viewed as functions of $\mathcal{P}$. As such, $\widehat{\mathcal{P}}$ is adaptively determined by minimizing the penalized least square function in (3).

To maximize the expected outcome of interest, our proposed I2DR is then given by

$$\widehat{d}(x)=\mathop{\arg\max}_{\mathcal{I}\in\widehat{\mathcal{P}}}\widehat{q}_{\mathcal{I}}(x),\qquad \forall x\in\mathbb{X}. \tag{4}$$

When the argmax in (4) is not unique, $\widehat{d}(\cdot)$ outputs the interval that contains the smallest treatment.

We next evaluate the value function under the proposed I2DR $V(\widehat{d})$ and $V(d^{opt})$. For each interval $\mathcal{I}$ in the estimated optimal partition $\widehat{\mathcal{P}}$, we estimate the generalized propensity score function $e(\mathcal{I}|x) \equiv \Pr(A \in \mathcal{I}|X = x)$. Let $\widehat{e}(\mathcal{I}|x)$ denote the resulting estimate. Following the estimation strategy in Zhang et al. (2012), we propose the following value estimator under (4),

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)}\big\{Y_i - \max_{\mathcal{I}\in\widehat{\mathcal{P}}}\widehat{q}_{\mathcal{I}}(X_i)\big\} + \max_{\mathcal{I}\in\widehat{\mathcal{P}}}\widehat{q}_{\mathcal{I}}(X_i)\right]. \tag{5}$$

Statistical properties of the estimates in (4) and (5) are studied in Section 4. Although we use the example of piecewise functions to motivate our procedure, the proposed method allows the outcome regression function to be a continuous function of a and x as well. Specifically, we can approximate any continuous outcome function by a piecewise function with the increasing number of partitions. As such, the proposed jump-interval learning method is applicable to handle Model II as well. See Section 4 for detail.

3.2 Linear- and Deep-JIL

In practice, we consider two concrete proposals for implementing jump interval-learning, by considering a linear function class and a deep neural networks (DNN) class for $\mathcal{Q}_{\mathcal{I}}$ in (3). In particular, the proposed L-JIL yields a set of linear decision rules that is easy to compute and interpret. See the real data analysis in Section 6 for details. In theory, it achieves a better convergence rate under the correct model specification. We recommend using L-JIL in applications where a simple and interpretable decision rule is preferred. On the contrary, the use of DNN in D-JIL allows us to capture the complicated outcome-covariates relationships that often occur in high-dimensional settings. We recommend using D-JIL in applications with high-dimensional covariates and complicated nonlinear associations.

We also remark that the proposed method is very general and allows a large variety of function approximators. Although we focus on linear models and deep neural nets in this paper, other function approximators such as linear basis expansion, reproducing kernel Hilbert spaces, and random forests are equally applicable. The theoretical properties of the resulting estimated Q-function can be similarly established (see e.g., Burman and Chen, 1989; Steinwart and Christmann, 2008; Wager and Athey, 2018, for the convergence rate of these nonparametric estimators).

3.2.1 Case 1: Linear-JIL

We use a linear regression model for $\mathcal{Q}_{\mathcal{I}}$. Specifically, we set $q(x, \theta_{\mathcal{I}})$ to $\bar{x}^\top\theta_{\mathcal{I}}$ for any interval $\mathcal{I}$ and $x \in \mathbb{X}$, where $\bar{x}$ is a shorthand for the vector $(1, x^\top)^\top$. Adopting the linearity assumption, we have $\widehat{q}_{\mathcal{I}}(x) = \bar{x}^\top\widehat{\theta}_{\mathcal{I}}$ for some $\widehat{\theta}_{\mathcal{I}}$. It follows from (4) that the proposed I2DR corresponds to a linear decision rule, i.e., $\widehat{d}(x) = \arg\max_{\mathcal{I}\in\widehat{\mathcal{P}}}\bar{x}^\top\widehat{\theta}_{\mathcal{I}}$. As such, the linearity assumption ensures our I2DR is interpretable to the domain experts.

We next discuss how to compute $\widehat{\mathcal{P}}$ and $\{\widehat{\theta}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{P}}\}$. The objective function in (3) is reduced to

$$\begin{aligned}&(\widehat{\mathcal{P}}, \{\widehat{\theta}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{P}}\}) = &(6)\\ &= \mathop{\arg\min}_{(\mathcal{P}\in\mathcal{B}(m),\{\theta_{\mathcal{I}}:\mathcal{I}\in\mathcal{P}\})}\left\{\sum_{\mathcal{I}\in\mathcal{P}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top\theta_{\mathcal{I}})^2 + \lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}\|_2^2\right) + \gamma_n|\mathcal{P}|\right\},\end{aligned}$$

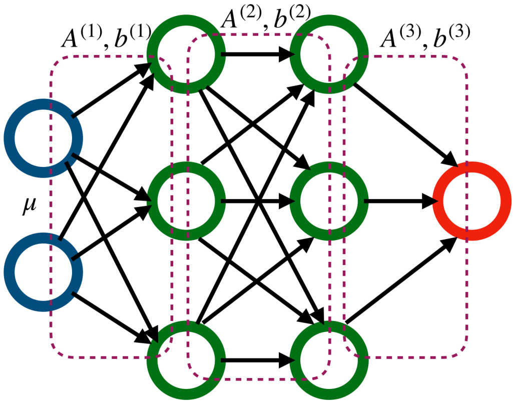

Figure 1: Illustration of DNN with $L = 2$ and $W = 25$; here $\mu \in \mathbb{R}^p$ is the input, the output is given by $A^{(3)}\sigma(A^{(2)}\sigma(A^{(1)}\mu + b^{(1)}) + b^{(2)}) + b^{(3)}$ where $A^{(l)}$, $b^{(l)}$ denote the corresponding parameters to produce the linear transformation for the $(l-1)$th layer and that σ denotes the componentwise rectified linear unit (ReLU) function. In this example, $W = \sum_{j=1}^{3}(\|A^{(3)}\|_0 + \|b^{(3)}\|_0) = 25$ where $\| \bullet \|_0$ denotes the number of nonzero elements in the vector or matrix.

where $\overline{X}_i = (1, X_i^\top)^\top$. We refer to this step as linear jump interval-learning (L-JIL). The ridge penalty $\lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}\|_2^2$ in (6) guarantees that for any interval $\mathcal{I} \in \widehat{\mathcal{P}}$, the parameter $\widehat{\theta}_{\mathcal{I}}$ is well defined even when $\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I}) < p+1$ such that the matrix $\sum_i \mathbb{I}(A_i \in \mathcal{I})\overline{X}_i\overline{X}_i^\top$ is not invertible. It also prevents over-fitting and yields more accurate estimates in high-dimensional settings.

3.2.2 Case 2: Deep-JIL

We next consider using deep neural networks (DNNs) to approximate the outcome regression function, so as to capture the complex dependence between the outcome and covariates. Specifically, the network consists of p input units (colored in blue in Figure 1), corresponding to the covariates X. The hidden units (colored in green) are grouped in a sequence of L layers. Each unit in the hidden layer is determined as a nonlinear transformation of a linear combination of the nodes from the previous layer. The total number of parameters in the network is denoted by W. See Figure 1 for an illustration. The parameters in DNNs can be solved using a stochastic gradient descent algorithm. In our implementation, we apply the Multi-layer Perceptron (MLP) regressor (Pedregosa et al., 2011) for parameter estimation. We refer to the resulting optimization as deep jump interval-learning (D-JIL).

Finally, we remark that alternative to our approach, one may directly apply DNN that takes the covariate-treatment pair (X, A) as the input to learn the outcome regression function. However, the resulting estimator for the outcome regression function is not guaranteed to be a piecewise function of the treatment. As such, it cannot yield an I2DR.

3.3 Implementation

In this section, we present the computational details for jump interval-learning. We employ the dynamic programming algorithm (see e.g., Friedrich et al., 2008) to find the optimal partition $\widehat{\mathcal{P}}$ that minimizes the objective function (3). Meanwhile, other algorithms for multi-scale change point detection are equally applicable (see e.g., Scott and Knott, 1974; Harchaoui and Lévy-Leduc, 2010; Fryzlewicz, 2014). Specifically, we adopt the PELT method proposed by Killick et al. (2012) that includes additional pruning steps within the dynamic programming framework to achieve a linear computational cost. Given $\widehat{\mathcal{P}}$, the set of functions $\{\widehat{q}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{P}}\}$ can be computed via either linear regression or deep neural network.

To detail our procedure, for any interval $\mathcal{I} \in [0,1]$, we define the cost function

$$\text{cost}(\mathcal{I}) = \min_{q(\cdot;\theta_{\mathcal{I}})\in\mathcal{Q}_{\mathcal{I}}} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})\big\{Y_i - q(X_i;\theta_{\mathcal{I}})\big\}^2 + \lambda_n ||\theta_{\mathcal{I}}||_2^2 \right],$$

where $\mathcal{Q}_{\mathcal{I}}$ is a class of linear functions or deep neural networks, corresponding to L-JIL and D-JIL, respectively.

For any integer $1 \le r < m$, denote by $\mathcal{B}(m,r)$ the set consisting of all possible partitions $\mathcal{P}_r$ of $[0, r/m)$ such that the end-points of each interval $\mathcal{I} \in \mathcal{P}_r$ lie on the grid $\{j/m : j = 0, 1, \ldots, r\}$. Set $\mathcal{B}(m,m) = \mathcal{B}(m)$, we define the Bellman function

$$B(r) = \inf_{\mathcal{P}_r \in \mathcal{B}(m,r)} \left(\sum_{\mathcal{I}\in\mathcal{P}_r} \text{cost}(\mathcal{I}) + \gamma_n(|\mathcal{P}_r| - 1) \right).$$

Let $B(0) = -\gamma_n$, the dynamic programming algorithm relies on the following recursion formula,

$$B(r) = \min_{j\in\mathcal{R}_r} \left\{ B(j) + \gamma_n + \text{cost}([j/m, r/m)) \right\}, \qquad \forall r \ge 1, \tag{7}$$

where $\mathcal{R}_r$ is the candidate change-point list updated by

$$\{j \in \mathcal{R}_{r-1} \cup \{r-1\} : B(j) + \text{cost}([j/m, (r-1)/m)) \le B(r-1)\}, \tag{8}$$

during each iteration with $\mathcal{R}_0 = \{0\}$. The constraint listed in (8) iteratively updates the set of candidate change points and removes values that can never be the minima of the objective function. It speeds up the computation, leading to a cost that is linear in the number of observations (Killick et al., 2012).

We briefly summarize our algorithm below. For a given integer r, we search the optimal change point location j that minimizes the above Bellman function $B(r)$ in (7). This requires applying the linear/MLP regression to learn $\widehat{q}_{[j/m, r/m)}$ and $\text{cost}([j/m, r/m))$ for each $j \in \mathcal{R}_r$. Let j^* be the corresponding minimizer. We then define the change points list $\tau(r) = \{j^*, \tau(j^*)\}$. This procedure is iterated to compute $B(r)$ and $\tau(r)$ for $r = 1, \ldots, m$. The optimal partition $\widehat{\mathcal{P}}$ is determined by the values stored in $\tau(\cdot)$. A pseudocode containing more details is given in Algorithm 1.

```
Global: data $\{(X_i, A_i, Y_i) : i = 1, \dots, n\}$; sample size $n$; covariates dimension $p$;
        number of initial intervals $m$; penalty terms $\lambda_n$, $\gamma_n$.
Local: integers $l, r \in \mathbb{N}$; cost dictionary $\mathcal{C}$; a vector of integers $\tau \in \mathbb{N}^m$;
       Bellman function $B \in \mathbb{R}^m$; a set of candidate point lists $\mathcal{R}$.
Output: $\widehat{\mathcal{P}}$ and $\{\widehat{q}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{P}}\}$.

I. Initialization. Set $B(0) \leftarrow -\gamma_n$; $\widehat{\mathcal{P}} \leftarrow Null$; $\tau \leftarrow Null$; $\mathcal{R}(0) \leftarrow \{0\}$;
II. Apply the PELT method. For $r = 1, \dots, m$:
    1. Compute $B(r) = \min_{j \in \mathcal{R}(r)} \{B(j) + \mathcal{C}([j/m, r/m)) + \gamma_n\}$ by Algorithm 2;
    2. $j^* \leftarrow \arg\min_{j \in \mathcal{R}(r)} \{B(j) + \mathcal{C}([j/m, r/m)) + \gamma_n\}$;
    3. $\tau(r) \leftarrow \{j^*, \tau(j^*)\}$;
    4. $\mathcal{R}(r) \leftarrow \{j \in \mathcal{R}(r-1) \cup \{r-1\} : B(j) + \mathcal{C}([j/m, (r-1)/m)) \leq B(r-1)\}$;
III. Get Partitions. $\tau^* \leftarrow \tau(m)$; $r \leftarrow m$; $l \leftarrow \tau^*[r]$; While $r > 0$:
    1. Let $\mathcal{I} = [l/m, r/m)$ if $r < m$ else $\mathcal{I} = [l/m, 1]$;
    2. $\widehat{\mathcal{P}} \leftarrow \widehat{\mathcal{P}} \cup \mathcal{I}$;
    3. $\widehat{q}_{\mathcal{I}}(\cdot) \leftarrow \arg\min_q \sum_i \mathbb{I}(A_i \in \mathcal{I}) \{Y_i - q(X_i)\}^2$;
    4. $r \leftarrow l$; $l \leftarrow \tau^*[r]$;
return $\widehat{\mathcal{P}}$ and $\{\widehat{q}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{P}}\}$.
```

Algorithm 1: Jump interval-learning.

```
Global: data $\{(X_i, A_i, Y_i) : i = 1, \dots, n\}$; cost dictionary $\mathcal{C}$; an interval $\mathcal{I}$;
        penalty term $\lambda_n$.
Output: $\mathcal{C}$.

If $\mathcal{C}(\mathcal{I}) == NULL$:
    (1). Apply linear/MLP regression:
         $\widehat{q}_{\mathcal{I}}(\cdot) \leftarrow \arg\min_q \sum_i \mathbb{I}(A_i \in \mathcal{I}) \{Y_i - q(X_i; \theta_{\mathcal{I}})\}^2 + n\lambda_n |\mathcal{I}| \|\theta_{\mathcal{I}}\|_2^2$;
    (2). Set the cost $\mathcal{C}(\mathcal{I})$ to the objective value;
return $\mathcal{C}$.
```

Algorithm 2: Calculation of the cost function.

3.3.1 Analysis of Computational Complexity

We analyze the computational complexity of the proposed methods in this section. The main computation lies in the dynamic programming algorithm to find the change points as well as the estimation of the outcome regression function in L- and D-JIL.

First, recall that we use the PELT method to implement the dynamic programming. It requires at least $\mathcal{O}(m)$ computing steps and at most $\mathcal{O}(m^2)$ steps (Friedrich et al., 2008). According to Theorem 3.2 in Killick et al. (2012), the expected computational cost is $\mathcal{O}(m)$.

Second, for each step in PELT, we need to train the DNN or linear regression model to calculate the cost function. Here, the complexity of training the linear regression is well known and equals $\mathcal{O}(np^2+p^3)$ with sample size n and feature dimension p. To the contrary, the complexity of training a DNN depends on the model architecture. Suppose we use a fully connected MLP with w width and d depth, and set the total number of epochs for training to e. Then the time complexity is given by $\mathcal{O}\{ne(d-1)w^2\}$[1].

To summarize, the expected computational complexities of the proposed linear- and deep-JIL are given by $\mathcal{O}\{m(np^2+p^3)\}$ and $\mathcal{O}\{mne(d-1)w^2\}$, respectively.

3.4 Tuning Parameters

Our proposal requires specifying the tuning parameters m, λ_n, and γ_n. We first discuss the choice of m. In practice, we recommend setting $m = n/c$ with some constant $c > 0$ such that m and n are of the same order. The choice of c represents a trade-off between the estimation bias and the computational cost. A small value of c would improve the estimation efficiency whereas a larger value of c saves the computation time. We recommend using the smallest possible c whenever the computation is affordable. In our numerical studies, we tried several different values of c and found the resulting estimated I2DRs have approximately the same value function as long as c is not too large. Thus, the proposed I2DR is not overly sensitive to the choice of this constant. Detailed empirical results can be found in Sections 5.3 and 6.

We next discuss the choices of λ_n and γ_n. The selection of these tuning parameters relies on the concrete proposal to approximate the outcome regression function. We elaborate below.

3.4.1 Tuning in L-JIL

For L-JIL, we choose γ_n and λ_n simultaneously via cross-validation. The theoretical requirements of γ_n and λ_n for L-JIL are imposed in the statement of Theorem 1. We further develop an algorithm that substantially reduces the computation complexity resulting from the use of cross-validation.

To be more specific, let $\Lambda_n = \{\lambda_n^{(1)}, \cdots, \lambda_n^{(H)}\}$ and $\Gamma_n = \{\gamma_n^{(1)}, \cdots, \gamma_n^{(J)}\}$ be the set of candidate tuning parameters. For a given integer K_0, we randomly split the data into K_0 equal-sized subgroups. Let $\mathbb{G}_k$ denote indices of the subsamples in the kth subgroup, for $k = 1, \cdots, K_0$. Let $\mathbb{G}_{-k}$ denote the complement of $\mathbb{G}_k$. For any $\lambda_n \in \Lambda_n$, $\gamma_n \in \Gamma_n$, $k \in \{1, \cdots, K_0\}$, let $(\widehat{\mathcal{P}}_{\lambda_n,\gamma_n,k}, \{\widehat{\theta}_{\mathcal{I},\gamma_n,\lambda_n,k} : \mathcal{I} \in \widehat{\mathcal{P}}_{\lambda_n,\gamma_n,k}\})$ denote the optimizer (6), computed

[1] See https://ai.stackexchange.com/questions/5728/what-is-the-time-complexity-for-training-a-neural-network-using-back-propagation/5730.

based on the data in $\mathbb{G}_{-k}$. We aim to choose γ_n and λ_n that minimizes

$$\frac{1}{n}\sum_{k=1}^{K_0}\sum_{i\in\mathbb{G}_k}\sum_{\mathcal{I}\in\widehat{\mathcal{P}}_{\lambda_n,\gamma_n,k}}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I},\gamma_n,\lambda_n,k}\}^2. \tag{9}$$

To solve (9), we remark that there is no need to apply Algorithm 1 $|\Lambda_n|\times|\Gamma_n|$ times to compute the minimizer of (6) over the set of candidate tuning parameters. We develop an algorithm to facilitate the computation. The key observation is that, for any interval $\mathcal{I}\subseteq[0,1]$ and $k\in\{1,\cdots,K_0\}$, the set of estimators $\{\widehat{\theta}_{\mathcal{I},\gamma_n,\lambda_n,k}:\gamma_n\in\Gamma_n,\lambda_n\in\Lambda_n\}$ can be obtained simultaneously over the set of candidate tuning parameters. This forms the basis of our algorithm. More details are provided in Section A of the supplementary article.

3.4.2 Tuning in D-JIL

As for D-JIL, we find that the MLP regressor is not overly sensitive to the choice of λ_n, so we set $\lambda_n=0$. The parameter γ_n is chosen based on cross-validation. The theoretical requirement of γ_n for D-JIL is imposed in the statement of Theorem 2. To implement the cross-validation, we randomly split the data into K_0 equal-sized subgroups, denoted by $\{(X_i,A_i,Y_i)\}_{i\in\mathbb{G}_1}$, $\{(X_i,A_i,Y_i)\}_{i\in\mathbb{G}_2}$, $\cdots$, $\{(X_i,A_i,Y_i)\}_{i\in\mathbb{G}_{K_0}}$, accordingly. For each γ_n and $k=1,\dots,K_0$, we compute the estimators $\widehat{\mathcal{P}}_{\gamma_n,k}$ and $\widehat{q}_{\mathcal{I},\gamma_n,k}(\cdot)$ based on the sub-dataset in $\mathbb{G}_{-k}$. Then we choose γ_n that minimizes

$$\frac{1}{n}\sum_{k=1}^{K_0}\sum_{i\in\mathbb{G}_k}\sum_{\mathcal{I}\in\widehat{\mathcal{P}}_{\gamma_n,k}}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I},\gamma_n,k}(X_i)\}^2.$$

We also remark that implementing deep neural networks involves some other tuning parameters, such as the learning rate, and the numbers of hidden nodes and hidden layers. In our implementation, we set them to the default values of the MLP regressor implementation (Pedregosa et al., 2011).

4. Theory

We establish the statistical properties of our proposed method in this section. As we have commented, we allow the outcome regression function to be either a piecewise or continuous function of the treatment. We first study the statistical properties of L-JIL and D-JIL when Model I holds, respectively. We next outline a procedure to construct a confidence interval for the value under the proposed I2DR and prove its validity. Finally, we investigate the properties of our proposed method when Model II holds. These theoretical results imply that our method will work when the outcome regression function is either a piecewise or a continuous function.

4.1 Properties When Model I Holds

4.1.1 Results for L-JIL

To establish the theoretical properties of the I2DR obtained by L-JIL, we first assume (1) holds with $q_{\mathcal{I},0}(x)=\bar{x}^\top\theta_{\mathcal{I},0}$ for any $x\in\mathbb{X}$ and $\mathcal{I}\in\mathcal{P}_0$. In other words, the outcome regression function $Q(x,a)$ is linear in x and piecewise constant in a. Without loss of generality, assume $\theta_{0,\mathcal{I}_1} \neq \theta_{0,\mathcal{I}_2}$ for any two adjacent intervals $\mathcal{I}_1, \mathcal{I}_2 \in \mathcal{P}_0$. This guarantees that the representation in (1) is unique. We write $a_n \asymp b_n$ for two sequences $\{a_n\}, \{b_n\}$ if there exists some universal constant $c \geq 1$ such that $c^{-1} b_n \leq a_n \leq c b_n$. Define $\theta_0(\cdot) = \sum_{\mathcal{I} \in \mathcal{P}_0} \theta_{\mathcal{I},0} \mathbb{I}(\cdot \in \mathcal{I})$. Giving $(\widehat{\mathcal{P}}, \{\widehat{\theta}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{P}}\})$, our estimator for the function $\theta_0(\cdot)$ is defined by

$$\widehat{\theta}(\cdot) = \sum_{\mathcal{I} \in \widehat{\mathcal{P}}} \widehat{\theta}_{\mathcal{I}} \mathbb{I}(\cdot \in \mathcal{I}). \tag{10}$$

This yields a piecewise constant approximation of $\theta_0(\cdot)$. We first study the theoretical properties of $\widehat{\theta}(\cdot)$. Toward that end, we need to impose the following condition on the probability tails of X and Y.

(A4) Suppose there exists some constant $\omega > 0$ such that $\sup_{a,j} \|X^{(j)}\|_{\psi_2|A=a} \leq \omega$ and $\sup_a \|Y\|_{\psi_2|A=a} \leq \omega$ almost surely, where $X^{(j)}$ denotes the jth element of X, and that for any random variable Z, $\|Z\|_{\psi_2|A=a}$ denotes the conditional Orlicz norm given that $A = a$, i.e.,

$$\|Z\|_{\psi_2|A=a} \triangleq \inf_{C>0} \left[\mathrm{E}\left\{ \exp\left(\frac{|Z|^2}{C^2} \right) \middle| A = a \right\} \leq 2 \right].$$

We remark that Condition (A4) is automatically satisfied when the covariates and the outcomes are bounded.

Theorem 1 *Assume (A1)-(A4) hold and* (1) *holds with* $q_{\mathcal{I},0}(x) = \bar{x}^\top \theta_{\mathcal{I},0}$. *Assume A has a bounded probability density function on* $[0,1]$. *Assume* $m \asymp n$, $\lambda_n = O(n^{-1}\log n)$, $\{\gamma_n\}_{n\in\mathbb{N}}$ *satisfies* $\gamma_n \to 0$ *and* $\gamma_n n / \log n \to \infty$. *Then, there exists some constant* $\bar{c} > 0$ *such that the following events hold with probability at least* $1 - O(n^{-2})$*:*

(i) $|\widehat{\mathcal{P}}| = |\mathcal{P}_0|$.

(ii) $\max_{\tau \in J(\mathcal{P}_0)} \min_{\hat{\tau} \in J(\widehat{\mathcal{P}})} |\hat{\tau} - \tau| \leq \bar{c} n^{-1} \log n$.

(iii) $\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da \leq \bar{c} n^{-1} \log n$. ■

In Theorem 1, results in (i) show the model selection consistency of our jump penalized estimator. Results in (ii) imply that the estimated change point locations converge at a rate of $O_p(n^{-1}\log n)$. In (iii), we derive an upper error bound for the integrated ℓ_2 loss of $\widehat{\theta}(\cdot)$. As discussed in the introduction, the derivation of Theorem 1 is nontrivial. A number of technical lemmas (see Lemma 1-4 in Section B.1) are established to prove Theorem 1. These results can be easily extended to study general covariate-based change point models.

We next establish the convergence rate of $V^{opt} - V(\widehat{d})$, where $V^{opt} = V(d^{opt})$. The quantity $V^{opt} - V(\widehat{d})$ represents the difference between the optimal value and the value under the proposed I2DR. The smaller the difference, the better the I2DR. Notice that $V^{opt} \geq V(d)$ for any I2DR $d(\cdot)$. It suffices to provide an upper bound for $V^{opt} - V(\widehat{d})$. We impose the following condition.

(A5.) Assume for any $\mathcal{I}_1, \mathcal{I}_2 \in \mathcal{P}_0$, there exist some constants $\gamma, \delta_0 > 0$ such that

$$\Pr(0 < |q_{\mathcal{I}_1,0}(X) - q_{\mathcal{I}_2,0}(X)| \leq t) = O(t^\gamma),$$

where the big-O term is uniform in $0 < t \leq \delta_0$.

Condition (A5) is commonly assumed in the literature to derive a sharp convergence rate for the value function under the estimated optimal IDR (Qian and Murphy, 2011; Luedtke and Van Der Laan, 2016; Shi et al., 2020). It is very similar to the margin condition (Tsybakov, 2004; Audibert and Tsybakov, 2007) used in the classification literature. This condition is automatically satisfied with $\gamma = 1$ when $q_{\mathcal{I},0}(X)$ has a bounded probability density function for any $\mathcal{I} \in \mathcal{P}_0$.

Theorem 2 *Assume the conditions in Theorem 1 are satisfied. Further, assume (A5) holds. Then, we have*

$$V^{opt} - V(\widehat{d}) \leq \bar{c}(n^{-1}\log n)^{(1+\gamma)/2} + \bar{c}n^{-1}\log n, \tag{11}$$

for some constant $\bar{c} > 0$*, with probability at least* $1 - O(n^{-2})$*.* ■

When (A5) holds with $\gamma = 1$, Theorem 2 suggests that $V(\widehat{d})$ converges to the optimal value at a rate of $O_p(n^{-1})$ up to some logarithmic factor. Notice that the events defined in Theorem 1 and 2 occur with probability at least $1 - O(n^{-2})$. Since $\sum_{n\geq 1} n^{-2} < +\infty$, an application of the Borel-Cantelli lemma implies that these events will occur for sufficiently large n almost surely.

4.1.2 Results for D-JIL

We study the theoretical properties of the proposed I2DR based on D-JIL when Model I is correct. Similar to the linear case, we assume $q_{\mathcal{I}_1,0} \neq q_{\mathcal{I}_2,0}$ for any two adjacent intervals $\mathcal{I}_1, \mathcal{I}_2 \in \mathcal{P}_0$. For any $\mathcal{I}$, we set the regression class $\mathcal{Q}_{\mathcal{I}}$ to a general class of feedforward architecture with $L_{\mathcal{I}}$ hidden layers, $W_{\mathcal{I}}$ many number of parameters, and ReLU activation function (Farrell et al., 2021).

To derive the theoretical properties of D-JIL, we assume the outcome regression function is a smooth function of the baseline covariates (see assumption (A6) below). Meanwhile, D-JIL is valid when $Q(x,a)$ is a nonsmooth function of x as well (see e.g., Imaizumi and Fukumizu, 2019). Specifically, define the class of β-smooth functions (also known as Hölder smooth functions with exponent β) as

$$\Phi(\beta, c) = \left\{ h : \sup_{\|\alpha\|_1 \leq \lfloor\beta\rfloor} \sup_{x \in \mathbb{X}} |D^\alpha h(x)| \leq c, \sup_{\|\alpha\|_1 = \lfloor\beta\rfloor} \sup_{\substack{x,y \in \mathbb{X} \\ x \neq y}} \frac{|D^\alpha h(x) - D^\alpha h(y)|}{\|x - y\|_2^{\beta - \lfloor\beta\rfloor}} \leq c \right\},$$

for some constant $c > 0$, where $\lfloor\beta\rfloor$ denotes the largest integer that is smaller than β and D^α denotes the differential operator D^α denote the differential operator:

$$D^\alpha h(x) = \frac{\partial^{\|\alpha\|_1} h(x)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

We introduce the following conditions.

(A6.) Suppose $Q(\bullet, a) \in \Phi(\beta, c)$, and $p(a|\bullet) \in \Phi(\beta, c)$ for any a.

(A7.) Functions $\{\widehat{q}_{\mathcal{I}}\}_{\mathcal{I} \in \widehat{\mathcal{P}}}$ are uniformly bounded.

Assumption (A7) ensures that the optimizer would not diverge in the ℓ_∞ sense. Similar assumptions are commonly imposed in the literature to derive the convergence rates of DNN estimators (see e.g., Farrell et al., 2021). Combining (A7) with (A6) allows us to derive the uniform rate of convergence for the class of DNN estimators $\{\widehat{q}_{\mathcal{I}}\}_{\mathcal{I}\in\widehat{\mathcal{P}}}$. The following theorem summarizes the theoretical properties of the proposed method via deep neural networks.

Theorem 3 *Assume (A1)-(A3), (A6), (A7) and Model I hold. Assume X and Y are bounded variables, and A has a bounded probability density function on $[0,1]$. Assume $m \asymp n$, $\{\gamma_n\}_{n\in\mathbb{N}}$ satisfies $\gamma_n \to 0$ and $\gamma_n \gg n^{-2\beta/(2\beta+p)}\log^8 n$. Then, there exist some constant $\bar{c} > 0$ and DNN classes $\{\mathcal{Q}_{\mathcal{I}} : \mathcal{I}\}$ with $L_{\mathcal{I}} \asymp \log(n|\mathcal{I}|)$ and $W_{\mathcal{I}} \asymp (n|\mathcal{I}|)^{p/(2\beta+p)}\log(n|\mathcal{I}|)$ such that the resulting D-JIL estimator computed by (3) satisfies*

(i) $|\widehat{\mathcal{P}}| = |\mathcal{P}_0|$;

(ii) $\max_{\tau\in J(\mathcal{P}_0)}\min_{\hat{\tau}\in J(\widehat{\mathcal{P}})}|\hat{\tau}-\tau| \le \bar{c}n^{-2\beta/(2\beta+p)}\log^8 n$;

(iii) $E|Q(X,A) - \sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathbb{I}(A\in\mathcal{I})\widehat{q}_{\mathcal{I}}(X)|^2 da \le \bar{c}n^{-2\beta/(2\beta+p)}\log^8 n$,

with probability at least $1 - O(n^{-2})$. ■

Theorem 3 establishes the properties of our method under settings where the $Q(x,a)$ is a piecewise function in the treatment. Results in (i) imply that D-JIL correctly identifies the number of change points. Results in (ii) imply that any change point in $\mathcal{P}_0$ can be consistently identified at a convergence rate of $O_p(n^{-2\beta/(2\beta+p)})$ up to some logarithmic factors. Notice that we use the piecewise function $\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathbb{I}(a\in\mathcal{I})\widehat{q}_{\mathcal{I}}(x)$ to approximate the outcome regression function. In (iii), we show our estimator for function $Q(X,A)$ converges at a rate of $O_p(n^{-2\beta/(2\beta+p)})$ up to some logarithmic factors. The theoretical choices of $L_{\mathcal{I}}$ and $W_{\mathcal{I}}$ in Theorem 3 are consistent with the literature of DNN estimators (Imaizumi and Fukumizu, 2019; Farrell et al., 2021). These DNN architectures ensure the convergence rate of our estimator for function $Q(X,A)$, which achieves the minimax-optimal nonparametric rate of convergence under (A6) (see e.g., Stone, 1982).

We next establish the convergence rate of $V^{opt} - V(\widehat{d})$ when Model I holds in the following theorem.

Theorem 4 *Assume the conditions in Theorem 3 are satisfied. Further, assume (A5) holds. Then, we have*

$$V(\widehat{d}) \ge V^{opt} - O(1)(n^{-\frac{2\beta}{2\beta+p}}\log^8 n + n^{-\frac{2\beta(1+\gamma)}{(2\beta+p)(2+\gamma)}}\log^{\frac{8+8\gamma}{2+\gamma}} n), \tag{12}$$

with probability at least $1 - O(n^{-2})$. ■

Theorem 4 suggests that $V(\widehat{d})$ converges to the optimal value at a rate of $O_p\{n^{-\frac{2\beta(1+\gamma)}{(2\beta+p)(2+\gamma)}}\}$ up to some logarithmic factors. This rate is slower than the rate ($O_p(n^{-1})$ up to some logarithmic factor) we obtained in Theorem 2 where we posit a parametric (linear) model. Suppose the condition $4\beta(1+\gamma) > (2\beta+p)(2+\gamma)$ holds, it follows that $V(\widehat{d}) = V^{opt} + o_p(n^{-1/2})$. This observation forms the basis of our inference procedure in Section 4.1.3. Here, the extra margin parameter γ in our results is introduced by (A5) to bound the bias due to the estimated decision rule $\widehat{d}$. If the margin parameter γ goes to infinity, we only require the smooth parameter $\beta > p/2$ to obtain $V(\widehat{d}) = V^{opt} + o_p(n^{-1/2})$. This condition ($\beta > p/2$) is commonly assumed in the literature on evaluating average treatment effects (see e.g., Chernozhukov et al., 2017; Farrell et al., 2021).

4.1.3 Evaluation of the Value Function

Suppose Model I holds. When L-JIL is used, it follows from Theorem 2 that $V(\widehat{d}) = V^{opt} + o_p(n^{-1/2})$. When D-JIL is used, if the smoothness parameter β (see (A6)) and the margin parameter γ (see (A5)) satisfy $4\beta(1+\gamma) > (2\beta+p)(2+\gamma)$, it follows from Theorem 4 that $V(\widehat{d}) = V^{opt} + o_p(n^{-1/2})$. In the following, we derive the asymptotic normality of $\sqrt{n}(\widehat{V} - V^{opt})$. By Slutsky's theorem, this implies that $\sqrt{n}\{\widehat{V} - V(\widehat{d})\}$ is asymptotically normal as well.

(A8.) $[\mathrm{E}\{\widehat{e}(\mathcal{I}|X) - e(\mathcal{I}|X)\}^2]^{1/2} = o(n^{-1/4})$ and that $\widehat{e}(\mathcal{I}; \bullet)$ belongs to the class of VC-type functions with VC-index upper bounded by $O(n^{1/2})$ (see e.g. Chernozhukov et al., 2017, for a detailed definition of the VC-type class), for any $\mathcal{I} \in \mathfrak{I}(m)$.

The first part of Assumption (A8) requires the generalized propensity score function to converge at certain rates. Similar assumptions are commonly imposed in the causal inference literature to derive the asymptotic distribution of the estimated average treatment effect (see e.g., Chernozhukov et al., 2017). The second part of (A8) essentially controls the model complexity of the estimator $\widehat{e}$. The more complicated $\widehat{e}$ is, the larger the VC index. Under (A6), we can show (A8) holds when DNN is used to model the generalized propensity score.

Theorem 5 *Assume (A8) holds and suppose functions $\{\widehat{e}_{\mathcal{I}}\}_{\mathcal{I}\in\widehat{\mathcal{P}}}$ are uniformly bounded away from zero. Further assume that for any $\mathcal{I}_1, \mathcal{I}_2 \in \mathcal{P}_0$ with $\mathcal{I}_1 \neq \mathcal{I}_2$, we have $Pr(q_{\mathcal{I}_1,0}(X) = q_{\mathcal{I}_2,0}(X)) = 0$.*

(i) Suppose conditions in Theorem 2 are satisfied. Then, under L-JIL, we have

$$\sqrt{n}(\widehat{V} - V^{opt}) \xrightarrow{d} N(0, \sigma_L^2),$$

for some $\sigma_L^2 > 0$.

(ii) Suppose conditions in Theorem 4 are satisfied with $4\beta(1+\gamma) > (2\beta+p)(2+\gamma)$. Then, under D-JIL, we have

$$\sqrt{n}(\widehat{V} - V^{opt}) \xrightarrow{d} N(0, \sigma_D^2),$$

for some $\sigma_D^2 > 0$. ■

We now introduce the estimator for the asymptotic variance σ_L^2 or σ_D^2, and derive a Wald-type $1-\alpha$ CI for V^{opt}. Since $V(\widehat{d}) = V^{opt} + o_p(n^{-1/2})$, the proposed CI also covers $V(\widehat{d})$ with probability tending to $1-\alpha$. We estimate σ_L^2 or σ_D^2 by

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)}\{Y_i - \max_{\mathcal{I}\in\widehat{\mathcal{P}}}\widehat{q}_{\mathcal{I}}(X_i)\} + \max_{\mathcal{I}\in\widehat{\mathcal{P}}}\widehat{q}_{\mathcal{I}}(X_i) - \widehat{V}\right]^2,$$

where $\{\widehat{q}_{\mathcal{I}}(\cdot)\}$ corresponds to the value estimations under L-JIL or D-JIL.

The corresponding $1-\alpha$ CI is given by $\widehat{V} \pm z_{\alpha/2}\widehat{\sigma}$, where $z_{\alpha/2}$ denotes the upper $\alpha/2$-th quantile of a standard normal distribution. Similar to Theorem 5, we can show that $\widehat{\sigma}$ is consistent. This shows the validity of our inference procedure.

4.2 Properties When Model II Holds

4.2.1 Properties of L-JIL under Varying Coefficient Model

We first consider the case when the outcome regression function can be represented by a varying coefficient model and investigate the theoretical properties of the proposed L-JIL. Specifically, suppose the true outcome regression function takes the following form

$$Q(x,a) = \bar{x}^\top \theta_0(a), \qquad \forall x \in \mathbb{X}, a \in [0,1], \tag{13}$$

where $\bar{x} = (1, x^\top)^\top$ and $\theta_0(\cdot)$ is some continuous $(p+1)$-dimensional function. That is, we assume the conditional mean of the outcome is a linear function of individuals' covariates for any treatment $a \in [0,1]$. Yet, the model is flexible in that $\theta_0(\cdot)$ is allowed to be an arbitrary continuous function of a with certain smoothness constraints. Models of this type belong to the class of varying coefficient models popularly applied in many scientific areas (see e.g., Fan and Zhang, 2008, for an overview).

Here, we consider the following class of Hölder continuous functions for $\theta_0(\cdot)$. Suppose there exist some constants $L > 0$, $0 < \alpha_0 \le 1$ such that $\theta_0(\cdot)$ satisfies

$$\sup_{a_1, a_2 \in [0,1]} \|\theta_0(a_1) - \theta_0(a_2)\|_2 \le L|a_1 - a_2|^{\alpha_0}. \tag{14}$$

We first sketch a few lines to see why our method works under (14). For a given integer $k > 0$, we define $\theta_k^*(\cdot)$ as

$$\theta_k^*(a) = \sum_{j=0}^{k-1} \theta_0\left(\frac{j+1/2}{k+1}\right) \mathbb{I}(j \le (k+1)a < j+1) + \theta_0\left(\frac{k+1/2}{k+1}\right) \mathbb{I}((k+1)a \ge k).$$

Apparently, $\theta_k^*(\cdot)$ has at most k change points. In addition, with some calculations, we can show that $\sup_{a\in[0,1]} \|\theta_k^*(a) - \theta_0(a)\|_2 \le 2^{-\alpha_0}(k+1)^{-\alpha_0} L$. Letting $k \to \infty$, it is immediate to see that $\theta_0(\cdot)$ can be uniformly approximated by a step function as the number of change points increases.

In Theorems 1 and 2, we have shown the proposed I2DR is consistent under the piecewise linear function assumption. Based on the above discussion, we expect that jump interval-learning also works when the model (13) holds. We formally establish the corresponding theoretical results in the following theorem.

Theorem 6 *Assume (A1)-(A4) and* (14) *hold. Assume A has a bounded probability on* $[0,1]$. *Assume $m \asymp n$, $\lambda_n = O(n^{-1}\log n)$, γ_n satisfies $\gamma_n \to 0$ and $\gamma_n \gg n^{-1}\log n$. Under the model* (13), *there exists some constant $\bar{c} > 0$ such that the following holds with probability at least $1 - O(n^{-2})$:*

$$\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da \le \bar{c}\gamma_n^{2\alpha_0/(1+2\alpha_0)}.$$

In addition, assume $\gamma_n \asymp (n^{-1}\log n)^{(1+2\alpha_0)/(1+4\alpha_0)}$. Then there exists some constant $\bar{c}^* > 0$

such that the following occurs with probability at least $1 - O(n^{-2})$ *that*

$$V^{opt} - V(\widehat{d}) \leq \bar{c}^*(n^{-1}\log n)^{\alpha_0/(1+4\alpha_0)}. \tag{15}$$

■

It is worth mentioning that with proper choice of γ_n, the integrated ℓ_2 loss of $\widehat{\theta}(\cdot)$ converges at a rate of $O_p(n^{-2\alpha_0/(1+2\alpha_0)})$ up to some logarithmic factor. The rate is slower compared to the results in Theorem 1, since $\theta_0(\cdot)$ is only "approximately" piecewise constant. When $\theta_0(\cdot)$ is Lipschitz continuous, it follows from (15) that the value under our proposed I2DR will converge to the optimal value function at a rate of $O_p(n^{-1/5}\log^{1/5} n)$.

4.2.2 Properties of D-JIL under the Continuous outcome regression function

We next consider the general case when the outcome regression function is specified by model II and study the theoretical properties of the proposed D-JIL. The following theorem proves the consistency of the proposed estimator.

Theorem 7 *Suppose Q is a continuous function of a and x. Assume (A1)-(A3) and (A6)-(A7) hold. Assume X and Y are bounded variables, and A has a bounded probability density function on* $[0, 1]$. *Assume $m \asymp n$ and $\{\gamma_n\}_{n\in\mathbb{N}}$ satisfies $\gamma_n \to 0$ and $\gamma_n \gg n^{-2\beta/(2\beta+p)}\log^8 n$. Then, there exist some DNN classes $\{\mathcal{Q}_\mathcal{I} : \mathcal{I}\}$ with $L_\mathcal{I} \asymp \log(n|\mathcal{I}|)$ and $W_\mathcal{I} \asymp (n|\mathcal{I}|)^{p/(2\beta+p)}\log(n|\mathcal{I}|)$ such that the resulting D-JIL estimator computed by* (3) *satisfies*

(i) $\max_{\mathcal{I}\in\widehat{\mathcal{P}}}\sup_{a\in\mathcal{I}} E|\widehat{q}_\mathcal{I}(X) - Q(X,a)|^2 = O_p(\gamma_n^{\frac{2\alpha_0}{1+2\alpha_0}}) + O_p\big((n\gamma_n)^{-\frac{2\beta}{2\beta+p}}\log^8 n\big)$ *where the expectation is taken with respect to the marginal distribution of X;*

(ii) Suppose $\gamma_n \sim n^{-1/\left(1+\frac{\beta(2\alpha_0+1)}{\alpha_0(2\beta+p)}\right)}$. Then $V^{opt} - V(\widehat{d}) = O_p\big(n^{-\alpha_0\beta/(4\alpha_0\beta+\alpha_0 p+\beta)}\log^4 n\big)$.

■

Theorem 7 establishes the properties of our method under settings where Q is continuous in a. Results in (i) imply that $\widehat{q}_\mathcal{I}(x)$ can be used to uniformly approximate $Q(x,a)$ for any $a \in \mathcal{I}$. The consistency of the value in (ii) thus follows.

Finally, we remark that the optimal I2DR is well-defined under Model I. When Model II holds, however, it remains unclear whether the optimal I2DR is uniquely defined or not. Nonetheless, as shown in Theorems 6 and 7, the value under the proposed I2DR converges to that under the optimal IDR. This implies that even when the optimal I2DR is not uniquely defined, our proposal is able to identify one of them asymptotically.

4.2.3 Evaluation of the Value Function

As discussed in Section 4.1.3, the validity of the proposed CI for the optimal value requires the outcome regression function to satisfy the piecewise model assumption. Under Model II, however, the value under the proposed I2DR might not be $n^{-1/2}$-consistent to the optimal value. As such, in theory, the proposed CI would fail to cover the optimal value. Nonetheless, when the preference function π^* is assigned according to the generalized propensity score, i.e.,

$$\pi^*(a; x, \mathcal{I}) = \frac{p(a|x)}{\int_{\mathcal{I}} p(a|x) da},$$

and other regularity conditions hold, our CI is able to cover the value under the proposed I2DR. We omit the theoretical results to save space.

5. Simulations

5.1 Confidence Interval for the Value

In this section, we focus on scenarios where the outcome regression function takes the form of Model I and examine the coverage probability of the proposed CI in Section 4.1.3. Simulated data are generated from the following model:

$$Y|X, A \sim N(Q(X, A), 1), \quad A|X \sim \text{Unif}[0, 1] \text{ and } X^{(1)}, X^{(2)}, \dots, X^{(p)} \overset{iid}{\sim} \text{Unif}[-1, 1],$$

where Unif$[a, b]$ denotes the uniform distribution on the interval $[a, b]$. We consider the following two scenarios with different choices of $Q(X, A)$.

Scenario 1:

$$Q(x, a) = \begin{cases} 1 + x^{(1)}, & a < 0.35, \\ x^{(1)} - x^{(2)}, & 0.35 \le a < 0.65, \\ 1 - x^{(2)}, & a \ge 0.65. \end{cases}$$

Under Scenario 1, the outcome regression function is piecewise constant as a function of a, and is linear as a function of x. Here, we have $J(\mathcal{P}_0) = \{0.35, 0.65\}$ and $|\mathcal{P}_0| = 3$. With some calculations, one can show that the optimal value V^{opt} equals 1.34.

Scenario 2:

$$Q(x, a) = \begin{cases} 1 + (x^{(1)})^3, & a < 0.35, \\ x^{(1)} - \log(1.5 + x^{(2)}), & 0.35 \le a < 0.65, \\ 1 - \sin(0.5\pi x^{(2)}), & a \ge 0.65. \end{cases}$$

Under Scenario 2, the outcome regression function is piecewise constant as a function of a, but is nonlinear as a function of x. The change points are $J(\mathcal{P}_0) = \{0.35, 0.65\}$ with $|\mathcal{P}_0| = 3$. The optimal value equals 1.35, based on Monte Carlo approximations.

For each scenario, we set $p = 4$ and consider three different choices of the sample size, corresponding to $n = 200, 400, 800$. We apply the proposed L-JIL and D-JIL to both scenarios. The detailed implementation is discussed in Section 3.3. We set $m = n/5$, $\lambda_n = 0$, $\gamma_n = 4n^{-1}\log(n)$, and construct the CI for V^{opt} based on the procedure described in Section 4.1.3. Reported in Table 1 are the estimated value function $\widehat{V}$ with its standard error $\widehat{\sigma}$, the empirical coverage probabilities of the proposed confidence interval for V^{opt}, and the number of estimated partitions $|\widehat{\mathcal{P}}|$, aggregated over 500 simulations. In addition, we include the integrated ℓ_2 loss of the estimated varying coefficient $\widehat{\theta}(\cdot)$ via L-JIL.

Based on the results, it is clear that the estimated value function approaches the optimal value as the sample size increases for both methods. For instance, when $n = 800$, L-JIL

Table 1: The estimated optimal value $\widehat{V}$ with its standard error, the empirical coverage probability of its associated confidence interval, and the averaged number of estimated partitions computed by the proposed L-JIL and D-JIL.

		Scenario 1, $p = 4$			Scenario 2, $p = 4$		
		$n = 200$	$n = 400$	$n = 800$	$n = 200$	$n = 400$	$n = 800$
Method	Optimal value V^{opt}	1.34			1.35		
L-JIL	Estimated optimal value $\widehat{V}$	1.436	1.383	1.340	1.400	1.351	1.421
	Mean of standard error $\widehat{\sigma}$	0.129	0.091	0.066	0.120	0.085	0.065
	Coverage probabilities(%)	89.80	93.20	95.60	92.40	94.60	95.00
	Number of partitions $\lvert\widehat{\mathcal{P}}\rvert$	2.97	3.01	3.00	2.19	2.81	3.01
	Integrated ℓ_2 loss of $\widehat{\theta}(\cdot)$	0.371	0.175	0.111	0.786	0.389	0.269
D-JIL	Estimated optimal value $\widehat{V}$	1.297	1.338	1.345	1.333	1.331	1.349
	Mean of standard error $\widehat{\sigma}$	0.160	0.108	0.060	0.166	0.102	0.060
	Coverage probabilities(%)	90.60	93.60	96.00	95.60	93.80	95.00
	Number of partitions $\lvert\widehat{\mathcal{P}}\rvert$	2.98	3.25	3.18	2.95	3.10	3.08

obtained an estimated value of 1.340 under Scenario 1 on average. D-JIL yields an average value of 1.349 under Scenario 2. These values are very close to the truths 1.34 and 1.35, respectively. The performance of our proposed L-JIL and D-JIL are comparable under Scenario 1. In addition, as the sample size increases, the coverage probability of the Wald-type CI approaches to the nominal level. This verifies our theoretical findings in Theorem 5. It is worth noting that the CI computed via L-JIL achieves the nominal coverage under Scenario 2 where the outcome regression function is nonlinear in x. We suspect this is due to that the optimal I2DR is close to a linear decision rule despite the nonlinearity of the outcome regression function.

Moreover, the averaged estimated number of partitions $|\widehat{\mathcal{P}}|$ is approximately 3 for all settings. This demonstrates the consistency of the estimated number of partitions in Theorems 1 and 3. In addition, the integrated ℓ_2 loss of the estimated varying coefficient computed via L-JIL converges to 0, as the sample size increases. For example, when $n = 800$, $\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da$ equals 0.111 for Scenario 1 and 0.269 for Scenario 2. These values are fairly small by noting that $\int_0^1 \|\theta_0(a)\|_2^2 da = 2$. Notice that $\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da$ decays at a rate that is approximately proportional to n^{-1}. This verifies our theoretical findings in Theorem 1.

5.2 Value Function under the Proposed I2DR

In this section, we consider more general settings and compare the proposed procedure with the existing state-of-the-art methods that output single-valued decision rules. Similar to Section 5.1, we generate the data from the following model:

$$Y|X, A \sim N(Q(X, A), 1), \quad A|X \sim \text{Unif}[0, 1] \quad \text{and} \quad X^{(1)}, X^{(2)}, \dots, X^{(p)} \overset{iid}{\sim} \text{Unif}[-1, 1].$$

In addition to Scenarios 1 and 2, we consider several other choices of the outcome regression function, allowing the working model assumption in Model I or Model II to be violated in

Table 2: Simulation scenarios.

scenarios	piecewise in a?	linear in x?	$J(\mathcal{P}_0)$	$\lvert\mathcal{P}_0\rvert$	optimal rule	optimal value
1	✓	✓	$\{0.35, 0.65\}$	3	$\arg\max_{I \in \mathcal{P}_0} q_{\mathcal{I}}(x)$	1.34
2	✓	×	$\{0.35, 0.65\}$	3	$\arg\max_{I \in \mathcal{P}_0} q_{\mathcal{I}}(x)$	1.35
3	✓	×	$\{0.25, 0.5, 0.75\}$	4	$\arg\max_{I \in \mathcal{P}_0} q_{\mathcal{I}}(x)$	0.76
4	×	✓	N.A.	N.A.	$0.5\mathbb{I}(\bar{x}^\top \theta < 0)$	1.28
5	×	×	N.A.	N.A.	$0.5 + 0.25(x^{(1)} + x^{(2)})$	8

some scenarios. Specifically, similar to Scenarios 1 to 2, the outcome regression function in Scenario 3 is a piecewise constant function of the treatment. As commented earlier, these scenarios are motivated by various applications such as dynamic pricing where the expected demand of a product has jump discontinuities as a function of the charged price. In Scenarios 4 and 5, however, the outcome regression function is continuous in the treatment. In particular, Scenario 4 is known as the varying coefficient model that has been widely applied in many scientific domains. Scenario 5 has been considered by Chen et al. (2016) for the personalized dose finding.

Scenario 3:

$$Q(x,a) = \begin{cases} \sqrt{x^{(1)}/2 + 0.5}, & a < 0.25, \\ \sin(2\pi x^{(2)}), & 0.25 \le a < 0.5, \\ 0.5 - (x^{(1)} + x^{(2)} - 0.75)^2, & 0.5 \le a < 0.75, \\ 0.5, & a \ge 0.75. \end{cases}$$

Scenario 4:

$$Q(x,a) = \bar{x}^\top \{2|a - 0.5|\theta^*\},$$

where $\theta^* = (1, 2, -2, 0_{p-2}^\top)^\top$. By setting $\theta_0(a) = 2|a - 0.5|\theta^*$, it is immediate to see that $Q(x,a) = \bar{x}^\top \theta_0(a)$ and satisfies the condition in (13).

Scenario 5:

$$Q(x,a) = 8 + 4x^{(1)} - 2x^{(2)} - 2x^{(3)} - 10(1 + 0.5x^{(1)} + 0.5x^{(2)} - 2a)^2.$$

We apply the proposed L-JIL and D-JIL to Scenarios 1-5 to estimate the optimal I2DR, with $p = 20$ and $n \in \{50, 100, 200, 400, 800\}$. The tuning parameters in JILs are specified according to Section 3.4. Here, we set $m = n/c$ with $c = 10$ to save computational costs. In Section 5.3, we report results with $c \in \{6, 8\}$ and find the values under the estimated I2DRs are very similar to those with $c = 10$.

To evaluate the proposed I2DRs, we compare its value function $V(\widehat{d})$ with the values under estimated optimal IDRs obtained by the linear outcome-weighted learning (L-O-L) and the nonlinear outcome-weighted learning based on the Gaussian kernel function (K-O-L) proposed by Chen et al. (2016), and the Q-learning method based on the linear regression (Q-Linear). To implement L-O-L and K-O-L, we fix the parameter $\phi_n = 0.1$, and select other tuning parameters by five-fold cross-validation, as in Chen et al. (2016). Finally, to implement Q-Linear, we first fit the outcome on $\{X, X \times X, A, A^2, XA, X \times XA, XA^2, X \times XA^2\}$ via the linear regression where $X \times X$ means the quadratic and cross terms among

Table 3: The value function under the proposed I2DR and IDRs estimated based on outcome-weighted learning (L-O-L and K-O-L) and Q-learning with the linear regression (Q-Linear) for Scenarios 1-5.

	n	50	100	200	400	800
Scenario 1	L-JIL	0.783(0.016)	0.832(0.016)	1.080(0.014)	1.259(0.002)	1.297(0.001)
$V = 1.34$	D-JIL	0.914(0.012)	0.967(0.008)	1.050(0.005)	1.071(0.005)	1.138(0.001)
$p = 20$	L-O-L	0.558(0.004)	0.574(0.004)	0.600(0.005)	0.597(0.005)	0.583(0.005)
	K-O-L	0.335(0.008)	0.415(0.006)	0.441(0.006)	0.457(0.005)	0.489(0.004)
	Q-Linear	1.026(0.041)	1.055(0.038)	1.080(0.038)	1.048(0.029)	0.829(0.028)
Scenario 2	L-JIL	0.741(0.021)	0.854(0.020)	1.180(0.007)	1.266(0.001)	1.299(0.001)
$V = 1.35$	D-JIL	0.900(0.012)	0.978(0.008)	1.074(0.004)	1.102(0.003)	1.141(0.001)
$p = 20$	L-O-L	0.450(0.009)	0.448(0.006)	0.447(0.005)	0.429(0.004)	0.410(0.003)
	K-O-L	0.115(0.019)	0.213(0.010)	0.229(0.007)	0.241(0.004)	0.276(0.002)
	Q-Linear	1.048(0.039)	1.071(0.037)	1.080(0.036)	1.042(0.027)	0.772(0.034)
Scenario 3	L-JIL	0.227(0.020)	0.268(0.013)	0.372(0.008)	0.432(0.003)	0.511(0.002)
$V = 0.76$	D-JIL	0.453(0.019)	0.469(0.009)	0.511(0.005)	0.526(0.004)	0.545(0.002)
$p = 20$	L-O-L	0.002(0.010)	-0.009(0.008)	-0.060(0.006)	-0.090(0.005)	-0.107(0.004)
	K-O-L	-0.268(0.026)	-0.233(0.015)	-0.260(0.009)	-0.251(0.006)	-0.233(0.003)
	Q-Linear	0.601(0.039)	0.604(0.032)	0.597(0.022)	0.575(0.015)	0.315(0.032)
Scenario 4	L-JIL	0.553(0.013)	0.564(0.011)	0.630(0.011)	0.806(0.006)	0.882(0.002)
$V = 1.28$	D-JIL	0.612(0.014)	0.651(0.008)	0.684(0.004)	0.653(0.006)	0.801(0.001)
$p = 20$	L-O-L	0.525(0.016)	0.458(0.010)	0.375(0.004)	0.300(0.002)	0.237(0.001)
	K-O-L	0.236(0.007)	0.260(0.004)	0.252(0.003)	0.244(0.001)	0.246(0.001)
	Q-Linear	0.995(0.025)	0.995(0.026)	0.999(0.021)	0.998(0.025)	0.832(0.044)
Scenario 5	L-JIL	5.82(0.05)	6.41(0.02)	6.80(0.01)	7.02(0.01)	7.16(0.01)
$V = 8.00$	D-JIL	5.57(0.06)	5.79(0.03)	5.97(0.02)	6.10(0.01)	6.26(0.01)
$p = 20$	L-O-L	5.92(0.07)	6.75(0.03)	7.32(0.02)	7.66(0.01)	7.81(0.01)
	K-O-L	6.70(0.02)	7.05(0.02)	7.38(0.01)	7.58(0.01)	7.56(0.01)
	Q-Linear	-0.53(1.27)	1.35(1.05)	3.80(0.57)	6.57(0.21)	6.57(0.21)

X. Denote the resulting estimator as $\widehat{Q}^L(x,a)$, then the optimal dose for a patient with covariates $X = x$ is given by $\arg\max_a \widehat{Q}^L(x,a)$. All the value functions are evaluated via Monte Carlo simulations. The average value function as well as its standard deviation over 200 replicates are summarized in Table 3.

It can be seen from Table 2 that both L-JIL and D-JIL are very efficient when Model I (Scenarios 1-3) holds, and perform reasonably well when Model II (Scenario 4 and 5) holds or the sample size is small. For instance, the proposed L-JIL achieves a value of 1.297 in Scenario 1 and 1.299 in Scenario 2, when $n = 800$. These values are very close to the optimal values, given by 1.34 and 1.35. In addition, due to the largely increased feature dimension, extra noises compromise the performance of D-JIL in both Scenarios 1 and 2, by comparing the results in Table 1 with that in Table 2. Yet, in Scenario 3, the nonlinear setting is hard to be modeled by the linear pattern, and thus the proposed D-JIL performs consistently better than L-JIL, due to the capacity of deep neural networks in approximating complicated non-linear relationships. Moreover, the value of the proposed

I2DR increases with the sample size in most cases. This supports our theoretical findings in Section 4.

In comparison, the value function under the estimated IDR using L-O-L and K-O-L is no more than half of the optimal value for each setting in Scenarios 1 to 4. This is owing to the ‘V-structured’ non-linear complex optimal decision rule in Scenario 4, which violates the assumption in Chen et al. (2016) that the decision rules should be smooth over the entire space of the treatment. While our method still works well for such a varying coefficient model. This supports our theoretical findings in Theorem 6. In Scenario 5, L-O-L and K-O-L have better performance, as the true optimal decision rule is linear and the outcome regression function is very sensitive to the change of the treatment level a (by noticing that the coefficient of the quadratic term in Scenario 5 is 10). Our methods perform worse in this scenario. However, the value difference is not large. In addition, under Scenarios 1-3, both L-JIL and D-JIL achieve larger value functions than Q-Linear when $n = 800$, since the linear model misspecifies the true conditional mean function under Model I. Under Scenarios 4-5, Q-Linear performs reasonably and comparably well as our proposed methods, because Scenario 4 is linear in X and the conditional mean outcome function under Scenario 5 is correctly specified by Q-Linear. Yet, due to the largely increased feature dimension p, the input dimension in Q-Linear is $3p(p-1)/2 + 3p + 2$, which compromises the performance of Q-Linear under Scenario 5 when n is small. Among competing methods, the Q-Linear method has better performance than outcome-weighted learning methods by Chen et al. (2016) in Scenarios 1-4. Yet, all results of values based on Q-Linear have considerably larger variances than other methods. More importantly, our methods are able to derive the I2DR, which is more interpretable and easier to implement in practice.

In addition, we notice that D-JIL performs comparably to L-JIL in Scenario 1. This is because the DNN model with the ReLU activation function contains the linear model as a special case. Yet, the asymptotic rate of convergence of D-JIL is slower than that of L-JIL due to its complexity. When $n \geq 200$, D-JIL performs worse than L-JIL, as expected. In Scenario 2, although the outcome regression function is nonlinear in x, the resulting I2DR can be well-approximated by a linear decision rule. As such, L-JIL and D-JIL achieve similar performance.

Finally, we report the computation time of D-JIL in Table 4. The computing infrastructure used is a virtual machine containing the second generation Intel Xeon Scalable Processors with 16 processor cores and 64GB memory in the AWS Platform. It can be seen that the computation time increases approximately linearly with the sample size. The main computation lies in adaptively selecting γ_n via cross-validation. We remark that parallel computing can be employed to further reduce the computation time.

5.3 Choice of m

Recall that we set $m = n/c$ for some constant $c > 0$. In Section 5.2, we report our simulation results with $c = 10$. In this section, we set $c \in \{6, 8\}$ and report the corresponding results in Tables 5-7 to investigate the sensitivity of the proposed methods to the choice of c. We also include results with $c = 10$ for completeness. It can be seen from Tables 5 and 6 that the value functions are very similar across different choices of c. In addition, it can be seen from Table 7 that the averaged number of estimated intervals for the proposed I2DR is very

Table 4: The computation time (in minutes) of the proposed D-JIL.

	50	100	200	400	800
Scenario 1	0.90(0.04)	1.95(0.03)	4.96(0.05)	14.04(0.12)	35.48(0.21)
Scenario 2	0.78(0.02)	1.56(0.02)	4.07(0.04)	13.53(0.07)	35.46(0.12)
Scenario 3	0.67(0.01)	1.02(0.02)	2.50(0.04)	7.34(0.04)	23.20(0.06)
Scenario 4	0.70(0.01)	1.09(0.02)	3.72(0.04)	10.25(0.06)	15.80(0.19)
Scenario 5	1.01(0.01)	1.51(0.01)	2.30(0.02)	5.37(0.02)	16.32(0.05)

Table 5: The value function of the proposed I2DR under L-JIL for Scenarios 1-5 with different choices of $m = n/c$.

	n	50	100	200	400	800
Scenario 1	$c = 6$	0.813(0.019)	0.858(0.017)	1.027(0.014)	1.249(0.003)	1.289(0.001)
$V = 1.34$	$c = 8$	0.836(0.022)	0.870(0.018)	1.024(0.014)	1.238(0.002)	1.295(0.001)
$p = 20$	$c = 10$	0.783(0.016)	0.832(0.016)	1.080(0.014)	1.259(0.002)	1.297(0.001)
Scenario 2	$c = 6$	0.804(0.025)	0.891(0.021)	1.132(0.008)	1.257(0.002)	1.290(0.001)
$V = 1.35$	$c = 8$	0.857(0.029)	0.935(0.021)	1.123(0.009)	1.241(0.002)	1.299(0.001)
$p = 20$	$c = 10$	0.741(0.021)	0.854(0.020)	1.180(0.007)	1.266(0.001)	1.299(0.001)
Scenario 3	$c = 6$	0.280(0.023)	0.310(0.014)	0.339(0.008)	0.422(0.003)	0.504(0.002)
$V = 0.76$	$c = 8$	0.229(0.019)	0.325(0.014)	0.326(0.008)	0.417(0.003)	0.512(0.002)
$p = 20$	$c = 10$	0.227(0.020)	0.268(0.013)	0.372(0.008)	0.432(0.003)	0.511(0.002)
Scenario 4	$c = 6$	0.565(0.015)	0.561(0.012)	0.639(0.011)	0.818(0.006)	0.884(0.002)
$V = 1.28$	$c = 8$	0.563(0.015)	0.564(0.012)	0.627(0.011)	0.810(0.006)	0.882(0.002)
$p = 20$	$c = 10$	0.553(0.013)	0.564(0.011)	0.630(0.011)	0.806(0.006)	0.882(0.002)
Scenario 5	$c = 6$	5.81(0.05)	6.38(0.02)	6.78(0.01)	6.99(0.01)	7.09(0.01)
$V = 8.00$	$c = 8$	5.82(0.05)	6.40(0.02)	6.78(0.01)	7.02(0.01)	7.12(0.01)
$p = 20$	$c = 10$	5.82(0.05)	6.41(0.02)	6.80(0.01)	7.02(0.01)	7.16(0.01)

close to the ground truth under Scenarios 1-3 where the underlying models are piecewise constant. Under Scenarios 4-5 however, the number of estimated intervals grows with the sample size, as expected. In all cases, the averaged number of estimated intervals is not overly sensitive to the choice of m. Finally, we report the computation time of the proposed D-JIL in Table 8. It can be seen that the computation time increases with m and n, as expected.

6. Real Data Analysis

In this section, we illustrate the empirical performance of our proposed method on real data from the International Warfarin Pharmacogenetics Consortium (Consortium, 2009). Warfarin is a medication that is commonly used for preventing blood clots such as thrombosis and thromboembolism. Its effect is evaluated by the international normalized ratio (INR), which is a measurement of the time it takes for the blood to clot, with an ideal number

Table 6: The value function of the proposed I2DR under D-JIL for Scenarios 1-5 with different choices of $m = n/c$.

	n	50	100	200	400	800
Scenario 1	$c = 6$	0.941(0.012)	0.972(0.008)	1.028(0.004)	1.065(0.004)	1.127(0.001)
$V = 1.34$	$c = 8$	0.973(0.016)	0.990(0.008)	1.030(0.004)	1.053(0.005)	1.136(0.001)
$p = 20$	$c = 10$	0.914(0.012)	0.967(0.008)	1.050(0.005)	1.071(0.005)	1.138(0.001)
Scenario 2	$c = 6$	0.943(0.013)	0.980(0.008)	1.037(0.004)	1.087(0.003)	1.129(0.001)
$V = 1.35$	$c = 8$	1.002(0.015)	1.012(0.008)	1.039(0.004)	1.076(0.003)	1.137(0.001)
$p = 20$	$c = 10$	0.900(0.012)	0.978(0.008)	1.074(0.004)	1.102(0.003)	1.141(0.001)
Scenario 3	$c = 6$	0.475(0.018)	0.480(0.009)	0.481(0.006)	0.493(0.004)	0.521(0.002)
$V = 0.76$	$c = 8$	0.416(0.019)	0.497(0.009)	0.493(0.006)	0.506(0.003)	0.532(0.002)
$p = 20$	$c = 10$	0.453(0.019)	0.469(0.009)	0.511(0.005)	0.526(0.004)	0.545(0.002)
Scenario 4	$c = 6$	0.624(0.014)	0.655(0.008)	0.686(0.004)	0.687(0.005)	0.801(0.001)
$V = 1.28$	$c = 8$	0.622(0.014)	0.651(0.008)	0.684(0.004)	0.676(0.005)	0.801(0.001)
$p = 20$	$c = 10$	0.612(0.014)	0.651(0.008)	0.684(0.004)	0.653(0.006)	0.801(0.001)
Scenario 5	$c = 6$	5.49(0.06)	5.69(0.03)	5.82(0.02)	5.97(0.01)	6.12(0.01)
$V = 8.00$	$c = 8$	5.58(0.05)	5.77(0.03)	5.91(0.02)	6.04(0.01)	6.20(0.01)
$p = 20$	$c = 10$	5.57(0.06)	5.79(0.03)	5.97(0.02)	6.10(0.01)	6.26(0.01)

Table 7: The averaged number of estimated intervals computed by L-JIL with different choices of $m = n/c$.

	n	50	100	200	400	800
Scenario 1	$c = 6$	2.04(0.17)	2.16(0.15)	2.60(0.09)	3.00(0.01)	3.00(0.00)
$\|\mathcal{P}_0\| = 3$	$c = 8$	1.98(0.15)	2.15(0.14)	2.47(0.07)	3.01(0.01)	3.00(0.00)
$p = 20$	$c = 10$	1.78(0.14)	1.95(0.13)	2.76(0.10)	3.00(0.00)	3.00(0.00)
Scenario 2	$c = 6$	2.38(0.18)	2.76(0.16)	3.17(0.09)	3.00(0.00)	3.00(0.00)
$\|\mathcal{P}_0\| = 3$	$c = 8$	2.38(0.15)	3.03(0.16)	3.02(0.04)	3.00(0.00)	3.00(0.00)
$p = 20$	$c = 10$	2.00(0.15)	2.64(0.14)	3.12(0.07)	3.00(0.00)	3.00(0.00)
Scenario 3	$c = 6$	2.63(0.20)	3.59(0.21)	3.65(0.17)	3.34(0.03)	3.76(0.02)
$\|\mathcal{P}_0\| = 4$	$c = 8$	2.26(0.19)	3.40(0.19)	3.41(0.14)	3.34(0.03)	3.78(0.02)
$p = 20$	$c = 10$	2.24(0.18)	3.01(0.18)	3.48(0.13)	3.32(0.03)	3.75(0.02)
Scenario 4	$c = 6$	1.68(0.15)	1.59(0.13)	1.86(0.07)	2.86(0.04)	3.12(0.01)
/	$c = 8$	1.61(0.14)	1.55(0.11)	1.79(0.07)	2.80(0.04)	3.15(0.02)
$p = 20$	$c = 10$	1.58(0.13)	1.62(0.13)	1.82(0.07)	2.79(0.04)	3.13(0.02)
Scenario 5	$c = 6$	4.78(0.17)	6.56(0.12)	9.56(0.08)	15.08(0.08)	25.71(0.07)
/	$c = 8$	4.29(0.13)	6.28(0.13)	9.30(0.09)	14.36(0.08)	23.91(0.08)
$p = 20$	$c = 10$	3.91(0.11)	5.96(0.11)	8.67(0.08)	13.62(0.08)	22.00(0.08)

of 2.5. High doses of Warfarin are more beneficial than its lower doses, but may lead to a high risk of bleeding as well. Proper dosing of Warfarin is thus of significant importance. Yet, this problem is particularly challenging due to the complex interactions between War-

Table 8: The computation time (in minutes) of D-JIL with different choices of $m = n/c$.

	n	50	100	200	400	800
Scenario 1	$c = 6$	0.69(0.03)	2.36(0.05)	10.38(0.10)	33.00(0.22)	87.56(0.36)
	$c = 8$	1.21(0.03)	2.23(0.05)	6.33(0.09)	19.96(0.16)	53.75(0.29)
$p = 20$	$c = 10$	0.90(0.04)	1.95(0.03)	4.96(0.05)	14.04(0.12)	35.48(0.21)
Scenario 2	$c = 6$	0.71(0.04)	2.35(0.06)	10.35(0.11)	32.08(0.21)	86.53(0.26)
	$c = 8$	1.02(0.02)	2.27(0.03)	5.79(0.08)	18.54(0.11)	54.36(0.24)
$p = 20$	$c = 10$	0.78(0.02)	1.56(0.02)	4.07(0.04)	13.53(0.07)	35.46(0.12)
Scenario 3	$c = 6$	0.71(0.02)	2.33(0.03)	6.12(0.05)	17.23(0.10)	52.96(0.24)
	$c = 8$	0.79(0.01)	1.59(0.02)	4.28(0.03)	9.61(0.06)	32.52(0.17)
$p = 20$	$c = 10$	0.67(0.01)	1.02(0.02)	2.50(0.04)	7.34(0.04)	23.20(0.06)
Scenario 4	$c = 6$	1.15(0.02)	2.67(0.05)	7.45(0.08)	21.38(0.19)	52.07(0.38)
	$c = 8$	0.88(0.02)	1.99(0.02)	4.92(0.04)	13.38(0.06)	32.77(0.19)
$p = 20$	$c = 10$	0.70(0.01)	1.09(0.02)	3.72(0.04)	10.25(0.06)	15.80(0.19)
Scenario 5	$c = 6$	1.24(0.02)	2.49(0.03)	4.47(0.04)	8.91(0.05)	28.52(0.09)
	$c = 8$	0.86(0.01)	1.37(0.01)	3.53(0.02)	7.39(0.03)	20.80(0.09)
$p = 20$	$c = 10$	1.01(0.01)	1.51(0.01)	2.30(0.02)	5.37(0.02)	16.32(0.05)

farin and many commonly used medications (Holbrook et al., 2005). Nonetheless, existing methods are not able to recommend an individualized interval-based dose rule for Warfarin.

To develop the optimal I2DR for Warfarin dosing, we use the dataset provided by the International Warfarin Pharmacogenetics (Consortium, 2009) for analysis. We choose 6 baseline covariates, including age, height, weight, gender, the VKORC1.AG genotype, and the VKORC1.AA genotype. This yields a total of 3848 with complete records of baseline information. Here, the VKORC1 genotype has been shown to play a particularly large role in response to Warfarin (Wadelius et al., 2005). The outcome is defined as the negative absolute distance between the INR after the treatment and the ideal number of 2.5, i.e, $Y = -|\text{INR} - 2.5|$. Thus, a larger outcome represents a better balance between preventing blood clots and the risk of bleeding, with the optimal value of 0. We use the min-max normalization to convert the range of the dose level A into $[0, 1]$.

To implement L-JIL and D-JIL, we set $c = 5$, i.e., $m = n/5$, and select γ_n and λ_n via cross-validation, as in Section 5.2. To further evaluate the empirical performance of the proposed I2DRs, we compare their values with the value under the IDR estimated by K-O-L. Specifically, we randomly select 70% of the data to compute the proposed I2DR and the IDR obtained by K-O-L, and evaluate their value functions using the remaining dataset. We then iterate this procedure 50 times to calculate the average value function. For each iteration, the value function is estimated based on the nonparametric estimator proposed by Zhu et al. (2020).

Specifically, let $\mathbb{G}_{test}$ denote observations in the testing dataset. For the IDR $\widetilde{d}$ computed by K-O-L, we consider the following nonparametric estimator for its value function,

$$\widetilde{V}(\widetilde{d}) = \int_x \frac{\sum_{i\in\mathbb{G}_{test}} Y_i K(h_x^{-1}(x - X_i))K(h_a^{-1}(\widetilde{d}(x) - A_i))}{\sum_{i\in\mathbb{G}_{test}} K(h_x^{-1}(x - X_i))K(h_a^{-1}(\widetilde{d}(x) - A_i))}\left\{\sum_{i\in\mathbb{G}_{test}} \frac{K(h_x^{-1}(x - X_i))}{|\mathbb{G}_{test}|h_x^p}\right\}dx,$$

Table 9: The averaged value and their standard deviation under four different choices of $\pi^*(\cdot; x, \mathcal{I})$ in the real data application.

Choice of $\pi^*(\cdot; x, \mathcal{I})$	Minimum Dose	Maximum Dose	Mid-point Dose	Uniformly Sample
Under L-JIL	-0.329(0.008)	-0.329(0.007)	-0.328(0.008)	-0.328(0.008)
Under D-JIL	-0.333(0.012)	-0.332(0.011)	-0.333(0.012)	-0.333(0.012)

Table 10: The averaged value with their standard deviation under L-JIL with different choices of c in real data analysis.

Choice of c	$c = 6$	$c = 8$	$c = 10$
Estimated Value	-0.329(0.009)	-0.329(0.009)	-0.328(0.008)

where $K(\cdot)$ denotes the Gaussian kernel function, and h_x and h_a are some bandwidth parameters. The tuning parameters h_x and h_a are chosen according to the numerical results in Section 5 of Zhu et al. (2020).

Notice that the value function under the proposed I2DR $\widehat{d}(\cdot)$ depends on the preference function π^*. To evaluate $\widehat{d}$, we consider multiple preference functions, including the maximum value, the minimum value, the mid-point value, and the value uniformly at random. In particular, when π^* is set to the uniform density function, we compute $\widetilde{V}^{\pi^*}(\widehat{d})$ as

$$\int_x \int_{\widehat{d}(x)} \frac{1}{|\widehat{d}(x)|} \frac{\sum_{i\in\mathbb{G}_{test}} Y_i K(h_x^{-1}(x-X_i))K(h_a^{-1}(d(x)-A_i))}{\sum_{i\in\mathbb{G}_{test}} K(h_x^{-1}(x-X_i))K(h_a^{-1}(d(x)-A_i))} \left\{ \sum_{i\in\mathbb{G}_{test}} \frac{K(h_x^{-1}(x-X_i))}{|\mathbb{G}_{test}|h_x^p} \right\} dadx.$$

Reported in Table 9 are the averaged values under the proposed I2DRs computed via L-JIL and D-JIL, with the aforementioned four choices of π^*, aggregated over 10 replications. It can be seen that our method is not sensitive to the choice of π^*.

Over 50 iterations, the average value functions of our proposed I2DRs computed by L-JIL and D-JIL are -0.332 and -0.331, larger than the value -0.344 of the IDR obtained by K-O-L. These values mean that the INR of patients following the recommended I2DR is around 0.33 far from the ideal amount of 2.5. In contrast, using K-O-L would lead to a greater departure from the ideal amount. We also set c to 6, 8, or 10 when employing L-JIL and report the average value functions and their standard deviations in Table 10. It can be seen that the performance is similar across difference choices of c. In addition, among the 50 iterations, $|\widehat{\mathcal{P}}|$ computed by L-JIL equals 3 for 40 iterations. Let $\widehat{\theta}_1$, $\widehat{\theta}_2$, and $\widehat{\theta}_3$ denote the corresponding regression coefficients associated with these three subintervals. We report the means and standard deviations of the estimated regression coefficients across these 40 iterations in Table 11. It can be seen that except for the intercept term associated with the first subinterval, the standard deviations of other parameters are fairly small. In the rest 10 iterations, $|\widehat{\mathcal{P}}|$ is either 2 or 4. As such, the change points and parameter estimates are relatively stable across 50 iterations.

Finally, we apply L-JIL to the entire data without sample-splitting to compute an I2DR and illustrate its interpretability. It turns out that L-JIL partitions $[0, 1]$ into three subintervals: $[0, 0.02)$, $[0.02, 0.17)$, and $[0.17, 1]$. We report these regression coefficients in Table

Table 11: The means and standard deviations of regression coefficients associated with the three subintervals computed by L-JIL over 40 iterations.

	Intercept	Age	Weight	Height	Gender	VKORC1.AG	VKORC1.AA
$\widehat{\theta}_1$	-1.289(0.841)	0.005(0.036)	0.003(0.005)	0.005(0.005)	-0.123(0.105)	-0.493(0.140)	-0.533(0.141)
$\widehat{\theta}_2$	-1.900(0.163)	0.021(0.006)	0.004(0.001)	0.008(0.001)	-0.203(0.021)	-0.024(0.027)	-0.125(0.026)
$\widehat{\theta}_3$	-0.450(0.063)	0.014(0.002)	0.001(0.001)	0.001(0.001)	-0.026(0.009)	-0.007(0.006)	-0.116(0.012)

Table 12: The regression coefficients associated with the three subintervals computed by applying L-JIL to the entire data without sample-splitting.

	Intercept	Age	Weight	Height	Gender	VKORC1.AG	VKORC1.AA
$\widehat{\theta}_1$	-1.229	0.013	0.003	0.004	-0.130	-0.466	-0.541
$\widehat{\theta}_2$	-2.008	0.021	0.004	0.008	-0.212	-0.030	-0.127
$\widehat{\theta}_3$	-0.518	0.015	0.001	0.001	-0.032	-0.001	-0.118

12. According to Table 12, the proposed I2DR based on L-JIL gives us a clear interpretation of the effect of baseline information on the dose assignment rule. For instance, patients whose genotype VKORC1 is AG or AA are more likely to receive low doses of Warfarin to prevent bleeding; older patients with larger weights shall be treated with higher dose levels. Future experiments are warranted to confirm these scientific findings. In Figure 2, we give a virtual representation of the proposed I2DR under L-JIL. Specifically, for each of the 3848 patients in the whole dataset, we plot a 3-dimensional vector $(\bar{x}^\top\widehat{\theta}_1, \bar{x}^\top\widehat{\theta}_2, \bar{x}^\top\widehat{\theta}_3)^\top$ based on his/her covariates x. Patients that are recommended to receive dose levels in $[0, 0.02)$, $[0.02, 0.17)$, and $[0.17, 1]$ are colored in blue, orange, and green, respectively. That is, we classify patients into three groups according to the recommended dose interval. It can be seen from Figure 2 that these three subgroups are well separated and have comparable sample sizes. In Figure 3, we further plot the histograms of the recommended treatments (uniformly randomly sampled from the computed I2DR) and the received treatments.

7. Discussions

7.1 Diverging Number of Change Points

When Model I is true, we assume $|\mathcal{P}_0|$ is fixed to simplify the results in Theorems 1, 2, 3, and 4. Our theoretical results can be generalized to the situation where $|\mathcal{P}_0|$ diverges with n as well. Take L-JIL as an example. Similar to Theorem 1, we can show that the ℓ_2 integrated loss satisfies $\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da = O_p(|\mathcal{P}_0| n^{-1} \log n)$. Compared to the results in Theorem 1, the convergence rate here is slower by a factor $|\mathcal{P}_0|$. In addition, $|\mathcal{P}_0| = o(n/\log n)$ is required to guarantee the consistency of $\widehat{\theta}$.

We next present more technical details. In the proof of Theorem 6 (see Section B.10 for details), we consider a more general framework and establish the ℓ_2 integrated loss of $\widehat{\theta}(\cdot)$

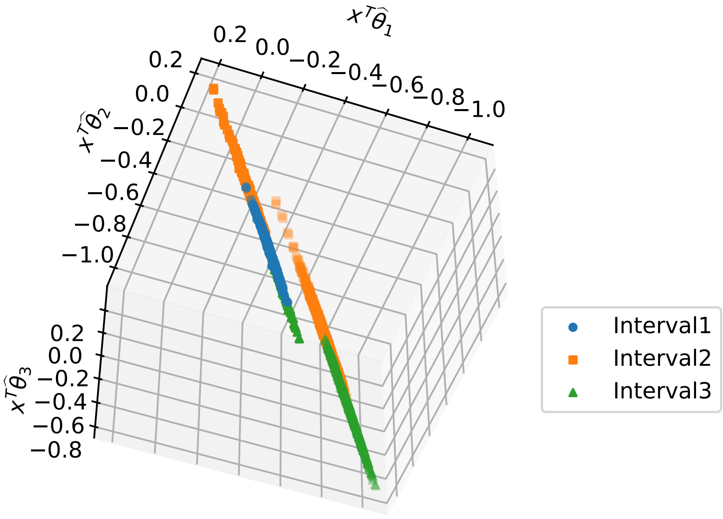

Figure 2: 3D plot of the proposed I2DR computed by L-JIL.

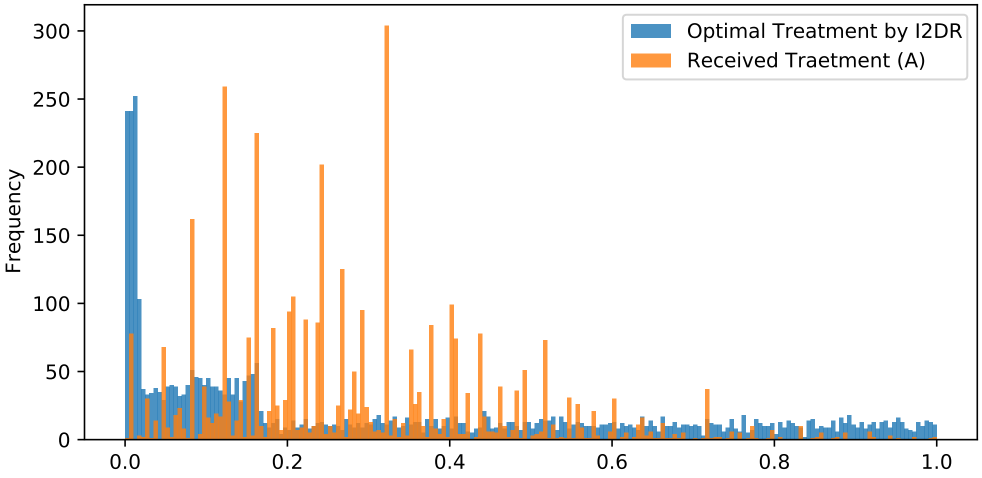

Figure 3: Histograms of the recommended treatments (uniformly randomly sampled from the computed I2DR) and the received treatments.

by assuming θ_0 satisfies $\limsup_{k\to\infty} k^{\alpha_0}\mathrm{AE}_k(\theta_0) < \infty$ where

$$\mathrm{AE}_k(\theta_0) = \inf_{\substack{\mathcal{P}:|\mathcal{P}|\le k+1\\ (\theta_{\mathcal{I}})_{\mathcal{I}\in\mathcal{P}}\in\prod_{\mathcal{I}\in\mathcal{P}}\mathbb{R}^{p+1}}} \left\{ \sup_{a\in[0,1]} \left\| \theta_0(a) - \sum_{\mathcal{I}\in\mathcal{P}} \theta_{\mathcal{I}}\mathbb{I}(a\in\mathcal{I}) \right\|_2 \right\}.$$

By definition, $\mathrm{AE}_k(\theta_0)$ describes how well $\theta_0(\cdot)$ can be approximated by a step function.

When $\theta_0(\cdot)$ is a step function with number of jumps equal to $|\mathcal{P}_0|$, we have $\mathrm{AE}_k(\theta_0) = 0$ for any $k \ge |\mathcal{P}_0|$. As a result, θ_0 satisfies the condition $\limsup_{k\to\infty} k^{\alpha_0}\mathrm{AE}_k(\theta_0) < \infty$ for any $\alpha_0 > 0$. As a result, the assertion (143) in the proof of Theorem 6 also holds for $\theta_0(\cdot)$ and we have with probability at least $1 - O(n^{-2})$ that

$$\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da \le O(1)(|\mathcal{P}_0|^{-\alpha_0} + \gamma_n|\mathcal{P}_0|),$$

where $O(1)$ denotes some positive constant. As $|\mathcal{P}_0| \to \infty$ and α_0 can be made arbitrarily large, we have with probability at least $1 - O(n^{-2})$ that

$$\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da \le O(1)(\gamma_n|\mathcal{P}_0|),$$

where $O(1)$ denotes some positive constant.

In Theorem 6, we require $\gamma_n \gg n^{-1}\log n$. However, this condition can be relaxed to $\gamma_n \ge \mathbb{M}_0 n^{-1}\log n$ for some sufficiently large constant $\mathbb{M}_0 > 0$. Under the latter condition, we have

$$\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da = O_p(|\mathcal{P}_0| n^{-1}\log n).$$

This yields the convergence rate of the ℓ_2 integrated loss of $\widehat{\theta}(\cdot)$.

7.2 Potential Alternative Approaches

In this paper, we focus on modeling the outcome regression function to derive I2DR. Below, we outline two other potential approaches and discuss their weaknesses.

7.2.1 A-learning Type Methods

Let's assume $q_{\mathcal{I}}(\cdot)$ satisfies (1) and the partition $\mathcal{P}_0$ is known to us. In order to eliminate the baseline function $u_0(\cdot)$, we can apply Robinson's transformation (see for example, Robinson, 1988; Chernozhukov et al., 2018; Zhao et al., 2022, and the references therein) and compute $\widetilde{q}_{\mathcal{I}}$ by minimizing

$$\operatorname*{arg\,min}_{\{q_{\mathcal{I}}\in\mathcal{Q}_{\mathcal{I}}:\mathcal{I}\in\mathcal{P}_0\}} \frac{1}{n}\sum_{i=1}^n [Y_i - \widehat{\mu}(X_i) - \sum_{\mathcal{I}\in\mathcal{P}_0} \{\mathbb{I}(A_i\in\mathcal{I}) - \widehat{e}(\mathcal{I}|X_i)\} q_{\mathcal{I}}(X)]^2,$$

where $\widehat{\mu}(x)$ correspond to some nonparametric estimators for $\mathrm{E}(Y|X=x)$. Both $\widehat{\mu}$ and $\widehat{e}$ can be obtained by some generic machine learning methods with good prediction performance.

When $\mathcal{P}_0$ is unknown, one might consider estimating $\mathcal{P}_0$ and $\{q_{\mathcal{I}} : \mathcal{I} \in \mathcal{P}_0\}$ jointly by

$$\operatorname*{arg\,min}_{\substack{\mathcal{P}\in\mathcal{B}(m),\\ \{q_{\mathcal{I}}\in\mathcal{Q}_{\mathcal{I}}:\mathcal{I}\in\mathcal{P}\}}} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \widehat{\mu}(X_i) - \sum_{\mathcal{I}\in\mathcal{P}}\{\mathbb{I}(A_i\in\mathcal{I}) - \widehat{e}(\mathcal{I}|X_i)\}q_{\mathcal{I}}(X_i)\right]^2 + \gamma_n|\mathcal{P}|,$$

for some tuning parameter γ_n. However, different from the objective function in (3), for a given partition $\mathcal{P}$, all the functions $\{q_{\mathcal{I}} : \mathcal{I} \in \mathcal{P}\}$ need to be jointed estimated. As a result, standard change point detection algorithms such as dynamic programming or binary segmentation (Scott and Knott, 1974) cannot be applied. The exhaustive search among all possible partitions is computationally infeasible. It remains unknown how to efficiently solve the above optimization problem. We leave it for future research.

7.2.2 Policy Search

As commented in Section 2.2, to apply value search, we need to specify a preference function π^*. To better illustrate the idea, let us suppose $\pi^*(\cdot; x, \mathcal{I}) = p(a|x)/\int_{x'\in\mathcal{I}} p(a|x')dx'$. That is, the preference function is the same as the one we observe in our data. Then, for a given I2DR d, we can consider the following inverse propensity score weighted estimator for $V^{\pi^*}(d)$,

$$\widehat{V}^{\pi^*}(d) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i\in d(X_i))}{\widehat{e}(d(X_i)|X_i)}Y_i.$$

For a given partition $\mathcal{P}$, let $\mathcal{D}_{\mathcal{P}}$ denote the space of I2DRs that we consider. Then $\widehat{d}$ can be computed by maximizing

$$\begin{aligned}
&\operatorname*{arg\,max}_{\mathcal{P}\in\mathcal{B}(m)}\operatorname*{arg\,max}_{d\in\mathcal{D}_{\mathcal{P}}}\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i\in d(X_i))}{\widehat{e}(d(X_i)|X_i)}Y_i\\
=\;&\operatorname*{arg\,max}_{\mathcal{P}\in\mathcal{B}(m)}\operatorname*{arg\,max}_{d\in\mathcal{D}_{\mathcal{P}}}\frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{I}\in\mathcal{P}}\frac{\mathbb{I}(A_i\in\mathcal{I})}{\widehat{e}(d(X_i)|X_i)}Y_i\mathbb{I}(d(X_i)=\mathcal{I}).
\end{aligned}$$

Suppose we consider the class of linear decision rules, i.e.,

$$\mathcal{D}_{\mathcal{P}} = \{d : d(x) = \operatorname*{arg\,max}_{\mathcal{I}\in\mathcal{P}} \theta_{\mathcal{I}}^\top \bar{x}\}.$$

It suffices to maximize

$$\operatorname*{arg\,max}_{\substack{\mathcal{P}\in\mathcal{B}(m)\\ \{q_{\mathcal{I}}\in\mathcal{Q}_{\mathcal{I}}:\mathcal{I}\in\mathcal{P}\}}} \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{I}\in\mathcal{P}}\frac{\mathbb{I}(A_i\in\mathcal{I})}{\widehat{e}(d(X_i)|X_i)}Y_i\mathbb{I}\{d(X_i) = \operatorname*{arg\,max}_{\mathcal{I}\in\mathcal{P}} q_{\mathcal{I}}(X_i)\}.$$

Similar to Section 7.2.1, for a given partition $\mathcal{P}$, all the functions $\{q_{\mathcal{I}} : \mathcal{I} \in \mathcal{P}\}$ need to be jointed estimated. As a result, dynamic programming cannot be applied. It remains unknown how to efficiently solve the above optimization problem. We leave it for future research.

7.3 Other Approaches

Recently, Meng et al. (2020) developed set-valued decision rules that contain equally beneficial treatments, borrowing ideas from multicategory classification with reject and refine options. Their method is developed under a discrete treatment setting with finitely many treatment options. It remains unclear whether it can be extended to our continuous treatment setting or not. We leave it for future research.

Next, in addition to the jump penalized regression formulation, one can alternatively consider the following constrained optimization function that directly restricts the number of estimated intervals to smaller than or equal to some integer M,

$$(\widehat{\mathcal{P}}, \{\widehat{\theta}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{P}}\}) = \operatorname*{arg\,min}_{(|\mathcal{P}| \le M, \{\theta_{\mathcal{I}}:\mathcal{I}\in\mathcal{P}\})} \sum_{\mathcal{I}\in\mathcal{P}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})(Y_i - q(X_i; \theta_{\mathcal{I}})^2 + \lambda_n |\mathcal{I}| \|\theta_{\mathcal{I}}\|_2^2 \right).$$

It would be interesting to further investigate the theoretical and numerical properties of this method. However, this is beyond the scope of the current paper and we leave it for future research.

Finally, our estimated partition is independent of the patient's baseline information. It might be practically more useful to consider patient-specific partitions and allow $\widehat{\mathcal{P}}$ to be a function of the baseline information. Additional interests include extending our work to policy evaluation (Cai et al., 2021) in the infinite horizon (Li et al., 2023). We leave them for future research.

7.4 Other Penalty Functions in L-JIL

In L-JIL, We use a ridge penalty in (6) to prevent overfitting in large p problems. When the true regression coefficient $\theta_0(\cdot)$ is sufficiently sparse, one can consider replacing the ridge penalty with the LASSO (Tibshirani, 1996) to improve the estimation accuracy. However, optimizing the resulting objective function requires computing the LASSO estimator $m(m-1)/2$ times. This is far more computationally expensive than the proposed method. It remains unknown whether the computation can be simplified. Finally, We leave it for future research.

Acknowledgments

The authors are very grateful to the anonymous reviewers for their valuable comments and suggestions. Chengchun Shi's research is partially supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/W014971/1.

References

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001217.

Leif Boysen, Angela Kempe, Volkmar Liebscher, Axel Munk, Olaf Wittich, et al. Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, 37(1):157–183, 2009.

Prabir Burman and Keh-Wei Chen. Nonparametric estimation of a regression function. *Ann. Statist.*, 17(4):1567–1596, 1989. ISSN 0090-5364. doi: 10.1214/aos/1176347382.

Hengrui Cai, Chengchun Shi, Rui Song, and Wenbin Lu. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34:15285–15300, 2021.

Bibhas Chakraborty, Susan Murphy, and Victor Strecher. Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.*, 19(3):317–343, 2010. ISSN 0962-2802. doi: 10.1177/0962280209105013.

Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516): 1509–1521, 2016.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato, et al. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econom. J.*, 21(1):C1–C68, 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097.

Victor Chernozhukov, Mert Demirer, Greg Lewis, and Vasilis Syrgkanis. Semi-parametric efficient policy learning with continuous actions. *Advances in Neural Information Processing Systems*, 32:15065–15075, 2019.

International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

Arnoud V den Boer and N Bora Keskin. Discontinuous demand functions: estimation and pricing. *Management Science*, 2020.

Caiyun Fan, Wenbin Lu, Rui Song, and Yong Zhou. Concordance-assisted learning for estimating optimal individualized treatment regimes. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(5):1565, 2017.

Jianqing Fan and Wenyang Zhang. Statistical methods with varying coefficient models. *Stat. Interface*, 1(1):179–195, 2008. ISSN 1938-7989. doi: 10.4310/SII.2008.v1.n1.a15.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

John M Flack and Bemi Adekola. Blood pressure and the new acc/aha hypertension guidelines. *Trends in cardiovascular medicine*, 30(3):160–164, 2020.

Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):495–580, 2014. ISSN 1369-7412. doi: 10.1111/rssb.12047. With 32 discussions by 47 authors and a rejoinder by the authors.

Felix Friedrich, Angela Kempe, Volkmar Liebscher, and Gerhard Winkler. Complexity penalized m-estimation: Fast computation. *Journal of Computational and Graphical Statistics*, 17(1):201–224, 2008.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, 42(6):2243–2281, 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1245.

Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.*, 105(492):1480–1493, 2010. ISSN 0162-1459. doi: 10.1198/jasa.2010.tm09181.

Anne M Holbrook, Jennifer A Pereira, Renee Labiris, Heather McDonald, James D Douketis, Mark Crowther, and Philip S Wells. Systematic overview of warfarin and its drug and food interactions. *Archives of internal medicine*, 165(10):1095–1106, 2005.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 869–878. PMLR, 2019.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251. PMLR, 2018.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107 (500):1590–1598, 2012.

Mariamma Kuruvilla and Cheryle Gurk-Turner. A review of warfarin dosing and monitoring. *Proceedings (Baylor University. Medical Center)*, 14(3):305, 2001.

Eric B Laber and Ying-Qi Zhao. Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514, 2015.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Yuhan Li, Wenzhuo Zhou, and Ruoqing Zhu. Quasi-optimal learning with continuous treatments. *International Conference on Learning Representations*, 2023.

Shuhan Liang, Wenbin Lu, Rui Song, and Lan Wang. Sparse concordance-assisted learning for optimal treatment decision. *The Journal of Machine Learning Research*, 18(1):7375–7400, 2017.

Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.

Haomiao Meng, Ying-Qi Zhao, Haoda Fu, and Xingye Qiao. Near-optimal individualized treatment recommendations. *Journal of Machine Learning Research*, 21(183):1–28, 2020.

S. A. Murphy. Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 65(2):331–366, 2003. ISSN 1369-7412. doi: 10.1111/1467-9868.00389. URL `https://doi.org/10.1111/1467-9868.00389`.

Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, pages 1–18, 2020.

Yue S Niu, Ning Hao, and Heping Zhang. Multiple change-point detection: A selective overview. *Statistical Science*, pages 611–623, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Zhengling Qi, Dacheng Liu, Haoda Fu, and Yufeng Liu. Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *Journal of the American Statistical Association*, 115(530):678–691, 2020.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Benjamin Rich, Erica EM Moodie, and David A Stephens. Simulating sequential multiple assignment randomized trials to generate optimal personalized warfarin dosing strategies. *Clinical trials*, 11(4):435–444, 2014.

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

P. M. Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. ISSN 0012-9682. doi: 10.2307/1912705.

JC Rotschafer, K Crossley, DE Zaske, K Mead, RJ Sawchuk, and LD Solem. Pharmacokinetics of vancomycin: observations in 28 patients and dosage recommendations. *Antimicrobial Agents and Chemotherapy*, 22(3):391–394, 1982.

Johannes Schmidt-Hieber et al. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

Juliana Schulz and Erica EM Moodie. Doubly robust estimation of optimal dosing strategies. *Journal of the American Statistical Association*, pages 1–13, 2020.

Andrew Jhon Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.

Jacob G Scott, Anders Berglund, Michael J Schell, Ivaylo Mihaylov, William J Fulp, Binglin Yue, Eric Welsh, Jimmy J Caudell, Kamran Ahmed, Tobin S Strom, et al. A genome-based model for adjusting radiotherapy dose (gard): a retrospective, cohort-based study. *The lancet oncology*, 18(2):202–211, 2017.

Chengchun Shi, Alin Fan, Rui Song, and Wenbin Lu. High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of statistics*, 46(3):925, 2018.

Chengchun Shi, Wenbin Lu, and Rui Song. Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21(176):1–67, 2020.

Rui Song, Weiwei Wang, Donglin Zeng, and Michael R Kosorok. Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 25(3):901, 2015.

Ingo Steinwart and Andreas Christmann. *Support vector machines.* Springer Science & Business Media, 2008.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, pages 1040–1053, 1982.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 58:267–288, 1996.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004. ISSN 0090-5364. doi: 10.1214/aos/1079120131.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. doi: 10.1007/978-1-4757-2545-2. With applications to statistics.

Mia Wadelius, LY Chen, K Downes, J Ghori, S Hunt, Niclas Eriksson, Ola Wallerman, Håkan Melhus, Claes Wadelius, D Bentley, et al. Common vkorc1 and ggcx polymorphisms associated with warfarin dose. *The pharmacogenomics journal*, 5(4):262–270, 2005.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Lan Wang, Yu Zhou, Rui Song, and Ben Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018.

C.J.C.H Watkins and P. Dayan. Q-learning. *Mach. Learn.*, 8:279–292, 1992.

Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2012.01763.x.

Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100 (3):681–694, 2013. ISSN 0006-3444. doi: 10.1093/biomet/ast014.

Yichi Zhang, Eric B. Laber, Anastasios Tsiatis, and Marie Davidian. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4):895–904, 2015. ISSN 0006-341X. doi: 10.1111/biom.12354.

Yichi Zhang, Eric B. Laber, Marie Davidian, and Anastasios A. Tsiatis. Estimation of optimal treatment regimes using lists. *J. Amer. Statist. Assoc.*, 113(524):1541–1549, 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1345743.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 84(2):382, 2022.

Ying-Qi Zhao, Donglin Zeng, Eric B. Laber, and Michael R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.*, 110(510):583–598, 2015. ISSN 0162-1459. doi: 10.1080/01621459.2014.937488.

Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.*, 107 (499):1106–1118, 2012. ISSN 0162-1459. doi: 10.1080/01621459.2012.695674.

Wenzhuo Zhou, Ruoqing Zhu, and Donglin Zeng. A parsimonious personalized dose-finding model via dimension reduction. *Biometrika*, 108(3):643–659, 2021.

Liangyu Zhu, Wenbin Lu, Michael R Kosorok, and Rui Song. Kernel assisted learning for personalized dose finding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 56–65, 2020.

Ruoqing Zhu, Ying-Qi Zhao, Guanhua Chen, Shuangge Ma, and Hongyu Zhao. Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, 73 (2):391–400, 2017.

This appendix is organized as follows. In Section A, we discuss more details on tuning parameters in L-JIL. Technical proofs are given in Section B.

Appendix A. More on Tuning in L-JIL

For L-JIL, we choose γ_n and λ_n simultaneously via cross-validation. As we will show below, the use of cross-validation will not increase the computation complexity substantially in L-JIL.

To elaborate, let us revisit the proposed jump interval-learning in Algorithm 1. The most time consuming part lies in computing the ridge-type estimator

$$\widehat{\theta}_{\mathcal{I}}(\lambda_n) = \left(\sum_{i \in \mathbb{G}_{-k}} \overline{X}_i \overline{X}_i^\top \mathbb{I}(A_i \in \mathcal{I}) + n\lambda_n |\mathcal{I}| \mathbb{E}_{p+1} \right)^{-1} \left(\sum_{i \in \mathbb{G}_{-k}} \overline{X}_i Y_i \mathbb{I}(A_i \in \mathcal{I}) \right), \tag{16}$$

where $\mathbb{E}_{p+1}$ is the identity matrix with dimension $p+1$, and the cost function

$$\text{cost}(\mathcal{I}, \lambda_n) = \frac{1}{n} \sum_{i \in \mathbb{G}_k} \mathbb{I}(A_i \in \mathcal{I}) \left\{ Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}}(\lambda_n) \right\}^2,$$

for any $\mathcal{I} \in \{[l/m, r/m) : 1 \le l < r < m\} \cup \{[l/m, 1] : 1 \le l < m\}$.

To compute $\{\widehat{\theta}_{\mathcal{I},\gamma_n,\lambda_n,k} : \gamma_n \in \Gamma_n, \lambda_n \in \Lambda_n\}$, we need to calculate $\{\widehat{\theta}_{\mathcal{I}}(\lambda_n) : \lambda_n \in \Lambda_n\}$ and $\{\text{cost}(\mathcal{I}, \lambda_n) : \lambda_n \in \Lambda_n\}$ for any $\mathcal{I}$. We first factorize the matrix $\sum_{i \in \mathbb{G}_{-k}} \overline{X}_i \overline{X}_i^\top \mathbb{I}(A_i \in \mathcal{I})$ as

$$\sum_{i \in \mathbb{G}_{-k}} \overline{X}_i \overline{X}_i^\top \mathbb{I}(A_i \in \mathcal{I}) = U \mathcal{T} U^\top,$$

according to the eigendecomposition, where U is some $(p+1) \times (p+1)$ orthogonal matrix and $\mathcal{T} = \text{diag}(\tau_0, \tau_1, \cdots, \tau_p)$ is some diagonal matrix. Let $\phi = U^\top \{\sum_{i \in \mathbb{G}_{-k}} \overline{X}_i Y_i \mathbb{I}(A_i \in \mathcal{I})\}$. Then the set of estimators $\{\widehat{\theta}_{\mathcal{I}}(\lambda_n) : \lambda \in \Lambda_n\}$ can be calculated by

$$\widehat{\theta}_{\mathcal{I}}(\lambda_n) = U \text{diag}\left\{ (\tau_0 + n\lambda_n |\mathcal{I}|)^{-1}, (\tau_1 + n\lambda_n |\mathcal{I}|)^{-1}, \cdots, (\tau_p + n\lambda_n |\mathcal{I}|)^{-1} \right\} \phi,$$

simultaneously for all λ_n.

Compared to separately inverting the matrix $\sum_{i \in \mathbb{G}_{-k}} \overline{X}_i \overline{X}_i^\top \mathbb{I}(A_i \in \mathcal{I}) + n\lambda_n |\mathcal{I}| \mathbb{E}_{p+1}$ in (16) for each λ_n to compute $\{\widehat{\theta}_{\mathcal{I}}(\lambda_n) : \lambda_n \in \Lambda_n\}$, the proposed method saves a lot of time especially for large p. Similarly, based on the eigendecomposition, we have

$$\begin{aligned} n\text{cost} \quad & (\mathcal{I}, \lambda_n) = \sum_{i \in \mathbb{G}_{-k}} Y_i^2 \mathbb{I}(A_i \in \mathcal{I}) \\ & -2\phi^\top \text{diag}\left\{ (\tau_0 + n\lambda_n |\mathcal{I}|)^{-1}, (\tau_1 + n\lambda_n |\mathcal{I}|)^{-1}, \cdots, (\tau_p + n\lambda_n |\mathcal{I}|)^{-1} \right\} \phi \\ & +\phi^\top \text{diag}\left\{ \tau_0(\tau_0 + n\lambda_n |\mathcal{I}|)^{-2}, \tau_1(\tau_1 + n\lambda_n |\mathcal{I}|)^{-2}, \cdots, \tau_p(\tau_p + n\lambda_n |\mathcal{I}|)^{-2} \right\} \phi, \end{aligned} \tag{17}$$

for all $\lambda_n \in \Lambda_n$. This facilitates the computation of $\{\text{cost}(\mathcal{I}, \lambda_n) : \lambda_n \in \Lambda_n\}$.

After obtaining these cost functions, we can recursively compute the Bellman function $B(r, \lambda_n, \gamma_n)$ by

$$B(r, \lambda_n, \gamma_n) = \min_{j \in \mathcal{R}_r} \left\{ B(j, \lambda_n, \gamma_n) + \gamma_n + \text{cost}([j/m, r/m), \lambda_n) \right\},$$

for all $r \geq 1$, $\lambda_n \in \Lambda_n$ and $\gamma_n \in \Gamma_n$. Given the Bellman function, the set of estimators $\{\widehat{\theta}_{\mathcal{I},\gamma_n,\lambda_n,k} : \gamma_n \in \Gamma_n, \lambda_n \in \Lambda_n\}$ thus can be computed efficiently.

Appendix B. Technical Proofs

In the proofs, we use $c, C > 0$ to denote some universal constants whose values are allowed to change from place to place. For any vector $\phi \in \mathbb{R}^q$, we use $\phi^{(j)}$ to denote the j-th element of ϕ, for any $j \in \{1, \dots, q\}$. For any two positive sequences $\{a_n\}$, $\{b_n\}$, $a_n \propto b_n$ means that $a_n \leq c b_n$ for some universal constant $c > 0$

B.1 Proof of Theorem 1

We provide the proof for Theorem 1 in this section. We present an outline of the proof first. Let $\delta_{\min} = \min_{\mathcal{I}\in\mathcal{P}_0} |\mathcal{I}|/3 > 0$. We divide the proof into four parts. In Part 1, we show that the following event occurs with probability at least $1 - O(n^{-2})$,

$$\max_{\tau\in J(\mathcal{P}_0)} \min_{\hat{\tau}\in J(\widehat{\mathcal{P}})} |\hat{\tau} - \tau| < \delta_{\min}. \tag{18}$$

By the definition of $\delta_{\min}$, this implies that

$$\Pr(|\widehat{\mathcal{P}}| \geq |\mathcal{P}_0|) \geq 1 - O(n^{-2}). \tag{19}$$

In Part 2, we show that

$$\max_{\tau\in J(\mathcal{P}_0)} \min_{\hat{\tau}\in J(\widehat{\mathcal{P}})} |\hat{\tau} - \tau| = O(n^{-1} \log n), \tag{20}$$

with probability at least $1 - O(n^{-2})$. This proves (ii) in Theorem 1. In Part 3, we prove

$$\Pr(|\widehat{\mathcal{P}}| \leq |\mathcal{P}_0|) \geq 1 - O(n^{-2}). \tag{21}$$

This together with (19) proves (i) in Theorem 1. In the last part, we show (iii) holds.

In the following, we first introduce some notations and auxiliary lemmas. Then, we present the proofs for Part 1, 2, 3 and 4.

Notations and technical lemmas: For any interval $\mathcal{I} \subseteq [0, 1]$, define

$$\begin{aligned}
\widehat{\theta}_{\mathcal{I}} &= \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})\overline{X}_i\overline{X}_i^\top + \lambda_n|\mathcal{I}|\mathbb{E}_{p+1}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})\overline{X}_i Y_i\right), \\
\theta_{0,\mathcal{I}} &= \left(\mathrm{E}\mathbb{I}(A \in \mathcal{I})\overline{X}\,\overline{X}^\top\right)^{-1} \{\mathrm{E}\mathbb{I}(A \in \mathcal{I})\overline{X}Y\},
\end{aligned}$$

where $\overline{X} = (1, X^\top)^\top$. It is immediate to see that the definition of $\widehat{\theta}_{\mathcal{I}}$ here is consistent with the one defined in (3) for any $\mathcal{I} \in \widehat{\mathcal{P}}$. In addition, under the model assumption in (13), the definition of $\theta_{0,\mathcal{I}}$ here is consistent with the one defined in step function model $\theta_0(a) = \sum_{\mathcal{I}\in\mathcal{P}_0} \theta_{0,\mathcal{I}}\mathbb{I}(a \in \mathcal{I})$ for any $\mathcal{I} \in \mathcal{P}_0$.

Let $\mathfrak{I}(m)$ denote the set of intervals

$$\begin{aligned}
\mathfrak{I}(m) \quad &= \quad \{[i_1/m, i_2/m) : \text{for some integers } i_1 \text{ and } i_2 \text{ that satisfy } 0 \leq i_1 < i_2 < m\} \\
&\cup \quad \{[i_3/m, 1] : \text{for some integers } i_3 \text{ that satisfy } 0 \leq i_3 < m\}.
\end{aligned}$$

Let $\{\tau_{0,k}\}_{k=1}^{K-1}$ with $0 < \tau_{0,1} < \tau_{0,2} < \cdots < \tau_{0,K-1} < 1$ be the locations of the true change points of $\theta_0(\cdot)$. Set $\tau_{0,0} = 0$, $\tau_{0,K} = 1$. We introducing the following lemmas.

Lemma 1 *Assume conditions in Theorem 1 are satisfied. Then there exist some constants $\bar{c}_0 > 0$, $c_0 \geq 1$ such that the following events occur with probability at least $1 - O(n^{-2})$: for any interval $\mathcal{I} \in \mathfrak{I}(m)$ that satisfies $|\mathcal{I}| \geq \bar{c}_0 n^{-1} \log n$, we have*

$$\|\widehat{\theta}_{\mathcal{I}} - \theta_{0,\mathcal{I}}\|_2 \leq \frac{c_0\sqrt{\log n}}{\sqrt{|\mathcal{I}|n}}, \tag{22}$$

$$\left\| \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})\overline{X}_i \right\|_2 \leq \frac{c_0\sqrt{|\mathcal{I}|\log n}}{\sqrt{n}}, \tag{23}$$

$$\left| \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \overline{X}^\top \theta_0(A_i)\}\overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\} \right| \leq \frac{c_0\sqrt{|\mathcal{I}|\log n}}{\sqrt{n}}, \tag{24}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})[\overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\}]^2 \geq \frac{1}{c_0}\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da - \frac{c_0\sqrt{|\mathcal{I}|\log n}}{\sqrt{n}}, \tag{25}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(|Y_i|^2 + \|\overline{X}_i\|_2^2) \leq c_0\left(\frac{\sqrt{|\mathcal{I}|\log n}}{\sqrt{n}} + |\mathcal{I}|\right). \tag{26}$$

In addition, for any $\mathcal{I} \in [0,1]$, we have

$$\|\theta_{0,\mathcal{I}}\|_2 \leq c_0. \tag{27}$$

■

Lemma 2 *Assume conditions in Theorem 1 are satisfied. Then there exist some constants $\bar{c}_1 > 0, c_1 \geq 1$ such that the following events occur with probability at least $1 - O(n^{-2})$: for any interval $\mathcal{I} \in \mathfrak{I}(m)$ that satisfies $\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \geq \bar{c}_1 n^{-1}\log n$,*

$$\begin{aligned} & \left| \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \overline{X}_i^\top \theta_0(A_i)\}\overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\} \right| \\ \leq \;& c_1\sqrt{n\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \log n}, \end{aligned} \tag{28}$$

$$\begin{aligned} & \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})[\overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\}]^2 \\ \geq \;& \frac{n}{c_1}\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da - c_1\sqrt{n\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \log n}. \end{aligned} \tag{29}$$

■

Lemma 3 *Assume conditions in Theorem 1 are satisfied. Then for sufficiently large n and any interval $\mathcal{I} \subseteq [0,1]$ of the form $[i_1, i_2)$ or $[i_1, i_2]$ with $i_2 = 1$ that satisfies $\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da = c_n$ for some sequence $\{c_n\}_n$ such that $c_n \geq 0, \forall n$ and $c_n \to 0$ as $n \to \infty$, we have either $\tau_{0,k-1} \leq i_1 \leq i_2 \leq \tau_{0,k}$ for some integer k such that $1 \leq k \leq K$ or*

$$\tau_{0,k-2} \leq i_1 < \tau_{0,k-1} < i_2 \leq \tau_{0,k} \quad \textit{and} \quad \min_{j\in\{1,2\}} |i_j - \tau_{0,k-1}| \leq c_2 c_n,$$

for some integer k such that $2 \le k \le K$ and some constant $c_2 > 0$, or

$$\tau_{0,k-3} \le i_1 < \tau_{0,k-2} < \tau_{0,k-1} < i_2 \le \tau_{0,k} \quad \textit{and} \quad \max_{j \in \{1,2\}} |i_j - \tau_{0,k-3+j}| \le \bar{c}_2 c_n,$$

for some integer k such that $3 \le k \le K$ and some constant $c_2 > 0$.

In addition, the following events occur with probability at least $1 - O(n^{-2})$: for any interval $\mathcal{I} \in \mathfrak{I}(m)$ that satisfies $\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \le \bar{c}_1 n^{-1} \log n$, we have

$$\left| \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \overline{X}_i^\top \theta_0(A_i)\} \overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\} \right| \le \bar{c}_2 \log n, \tag{30}$$

for some constant $\bar{c}_2 > 0$. ■

Lemma 4 *Under the conditions in Theorem 1, the following events occur with probability at least $1 - O(n^{-2})$: there exists some constant $\bar{c}_3 > 0$ such that $\min_{\mathcal{I} \in \widehat{\mathcal{P}}} |\mathcal{I}| \ge \bar{c}_3 \gamma_n$.* ■

Part 1: Assume $|\mathcal{P}_0| > 1$. Otherwise, (18) trivially hold. Consider the partition $\mathcal{P} = \{[0,1]\}$ which consists of a single interval and a zero vector $\mathbf{0}_{p+1}$. By definition, we have

$$\begin{aligned} &\sum_{\mathcal{I} \in \widehat{\mathcal{P}}} \left(\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + n\lambda_n |\mathcal{I}| \|\widehat{\theta}_{\mathcal{I}}\|_2^2 \right) + n\gamma_n |\widehat{\mathcal{P}}| \\ \le\ & \sum_{i=1}^n (Y_i - \overline{X}_i^\top \mathbf{0}_{p+1})^2 + n\lambda_n \|\mathbf{0}_{p+1}\|_2^2 + n\gamma_n = \sum_{i=1}^n Y_i^2 + n\gamma_n. \end{aligned}$$

In view of (26), we obtain with probability at least $1 - O(n^{-2})$,

$$\sum_{\mathcal{I} \in \widehat{\mathcal{P}}} \left(\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + n\lambda_n |\mathcal{I}| \|\widehat{\theta}_{\mathcal{I}}\|_2^2 \right) + n\gamma_n (|\widehat{\mathcal{P}}| - 1) \le c_0 n \left(\frac{\sqrt{\log n}}{\sqrt{n}} + 1 \right).$$

This implies that under the event defined in (26), we have for sufficiently large n,

$$\gamma_n (|\widehat{\mathcal{P}}| - 1) \le c_0 \left(\frac{\sqrt{\log n}}{\sqrt{n}} + 1 \right),$$

and hence

$$|\widehat{\mathcal{P}}| \le 2 c_0 \gamma_n^{-1}, \tag{31}$$

for sufficiently large n.

Under the event defined in Lemma 4, we have $\min_{\mathcal{I} \in \widehat{\mathcal{P}}} |\mathcal{I}| \ge \bar{c}_0 n^{-1} \log n$ for sufficiently large n, since $\gamma_n \gg n^{-1} \log n$. Thus, with probability at least $1 - O(n^{-2})$, the events defined in (22)-(26) hold for any interval $\mathcal{I} \in \widehat{\mathcal{P}}$.

Notice that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\widehat{\mathcal{P}}| \tag{32}$$

$$\geq \sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\gamma_n|\widehat{\mathcal{P}}| \geq n\gamma_n|\widehat{\mathcal{P}}|+\underbrace{\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{0,\mathcal{I}})^2}_{\eta_1}$$

$$+\underbrace{\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{\overline{X}_i^\top(\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}})\}^2}_{\eta_2}-\underbrace{2\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left|\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{0,\mathcal{I}})\overline{X}_i^\top(\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}})\right|}_{\eta_3}.$$

By (22) and (23), we obtain that

$$\begin{aligned}\eta_3 &\leq 2\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left\|\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{0,\mathcal{I}})\overline{X}_i\right\|_2\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2 \\ &\leq \sum_{\mathcal{I}\in\widehat{\mathcal{P}}}2c_0^2\log n\leq 2c_0^2\log n|\widehat{\mathcal{P}}|,\end{aligned} \tag{33}$$

with probability at least $1-O(n^{-2})$. Since $\gamma_n\gg n^{-1}\log n$, $\eta_2\geq 0$, for sufficiently large n, we have with probability at least $1-O(n^{-2})$,

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\widehat{\mathcal{P}}|\geq\eta_1. \tag{34}$$

Notice that

$$\eta_1=\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)+\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta_{0,\mathcal{I}}\}^2$$

$$=\underbrace{\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}^2}_{\eta_4}+\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta_{0,\mathcal{I}}\}^2$$

$$+\ 2\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\{\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta_{0,\mathcal{I}}\}.$$

Under the events defined in (24) and (25), it follows that

$$\begin{aligned}\eta_1 &\geq \eta_4+n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\frac{1}{c_0}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da-2c_0\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sqrt{|\mathcal{I}|n\log n} \\ &\geq \eta_4+n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\frac{1}{c_0}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da-2c_0\sqrt{|\widehat{\mathcal{P}}|n\log n},\end{aligned}$$

where the last inequality is due to Cauchy-Schwarz inequality. By (31) and the condition that $\gamma_n \gg n^{-1}\log n$, we obtain

$$\eta_1 \geq \eta_4 + n \sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \frac{1}{c_0}\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da + o(n), \tag{35}$$

with probability at least $1 - O(n^{-2})$. Notice that

$$\eta_4 = \sum_{i=1}^n \{Y_i - \overline{X}_i^\top \theta_0(A_i)\}^2 = \sum_{\mathcal{I}\in\mathcal{P}_0}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2.$$

Combining (34) with (35), we've shown that

$$\begin{aligned}
& \sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right) + n\gamma_n|\widehat{\mathcal{P}}| \\
\geq\ & \sum_{\mathcal{I}\in\mathcal{P}_0}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2 + n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\frac{1}{c_0}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2 da + o(n),
\end{aligned}$$

with probability at least $1 - O(n^{-2})$. By (27) and the condition that $\lambda_n = O(n^{-1}\log n)$, $\gamma_n = o(1)$, this further implies

$$\begin{aligned}
& \sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right) + n\gamma_n|\widehat{\mathcal{P}}| \\
\geq\ & \sum_{\mathcal{I}\in\mathcal{P}_0}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2 + n\lambda_n|\mathcal{I}|\|\theta_{0,\mathcal{I}}\|_2^2\right) + n\gamma_n|\mathcal{P}_0| \\
+\ & n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\frac{1}{c_0}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2 da + o(n).
\end{aligned} \tag{36}$$

For any integer k such that $1 \leq k \leq K-1$, let $\tau_{0,k}^*$ be the change point location that satisfies $\tau_{0,k}^* = i/m$ for some integer i and that $|\tau_{0,k} - \tau_{0,k}^*| < m^{-1}$. Denoted by $\mathcal{P}^*$ the oracle partition formed by the change point locations $\{\tau_{0,k}^*\}_{k=1}^{K-1}$. Set $\tau_{0,0}^* = 0$, $\tau_{0,K}^* = 1$ and $\theta^*_{[\tau^*_{0,k-1},\tau^*_{0,k})} = \theta_{0,[\tau_{0,k-1},\tau_{0,k})}$ for $1 \leq k \leq K-1$ and $\theta^*_{[\tau^*_{0,K-1},1]} = \theta_{0,[\tau_{0,K-1},1]}$. Let $\Delta_k = [\tau^*_{0,k-1},\tau^*_{0,k}) \cap [\tau_{0,k-1},\tau_{0,k})^c$ for $1 \leq k \leq K-1$ and $\Delta_K = [\tau^*_{0,K-1},1] \cap [\tau_{0,K-1},1]^c$. The length of each interval Δ_k is at most m^{-1}. Since $m \asymp n$, we have $m^{-1} \ll \bar{c}_0 n^{-1}\log n$. For any k and sufficiently large n, we can find an interval $\mathcal{I} \in \mathfrak{I}(m)$ with length between

$\bar{c}_0 n^{-1}\log n$ and $2\bar{c}_0 n^{-1}\log n$ that covers Δ_k. It follows that

$$\begin{aligned}
&\left\{\sum_{\mathcal{I}\in\mathcal{P}^*}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta^*_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\theta^*_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\mathcal{P}^*|\right\} \qquad (37)\\
-&\left\{\sum_{\mathcal{I}\in\mathcal{P}_0}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{0,\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\theta_{0,\mathcal{I}}\|_2^2\right)+n\gamma_n|\mathcal{P}_0|\right\}\\
\le\;& n\lambda_n\sup_{\mathcal{I}\subseteq[0,1]}\|\theta_{0,\mathcal{I}}\|_2^2+\sum_{k=1}^K\sum_{i=1}^n\mathbb{I}(A_i\in\Delta_k)\left(Y_i^2+\sup_{\mathcal{I}\subseteq[0,1]}\|\theta_{0,\mathcal{I}}\|_2^2\|\overline{X}_i\|_2^2\right)\\
\le\;& n\lambda_n\sup_{\mathcal{I}\subseteq[0,1]}\|\theta_{0,\mathcal{I}}\|_2^2+K\sup_{\substack{\mathcal{I}\in\mathfrak{I}(m)\\ 1\le\bar{c}_0^{-1}|\mathcal{I}|n\log^{-1}n\le 2}}\sum_{i=1}^n\mathbb{I}(A_i\in\mathcal{I})\left(Y_i^2+\sup_{\mathcal{I}\subseteq[0,1]}\|\theta_{0,\mathcal{I}}\|_2^2\|\overline{X}_i\|_2^2\right).
\end{aligned}$$

Since $\lambda_n=O(n^{-1}\log n)$, combining (37) together with (26) and (27), we obtain with probability at least $1-O(n^{-2})$,

$$\begin{aligned}
&\left\{\sum_{\mathcal{I}\in\mathcal{P}^*}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta^*_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\theta^*_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\mathcal{P}^*|\right\}\\
-&\left\{\sum_{\mathcal{I}\in\mathcal{P}_0}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{0,\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\theta_{0,\mathcal{I}}\|_2^2\right)+n\gamma_n|\mathcal{P}_0|\right\}\\
\le\;& c_0^2 n\lambda_n+K(c_0^2+1)c_0(\sqrt{2\bar{c}_0}+2\bar{c}_0)\log n=O(\log n)=o(n). \qquad (38)
\end{aligned}$$

By definition, we have

$$\begin{aligned}
&\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\widehat{\mathcal{P}}|\\
\le\;&\sum_{\mathcal{I}\in\mathcal{P}^*}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta^*_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\theta^*_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\mathcal{P}^*|.
\end{aligned}$$

In view of (36) and (38), we obtain that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da=o(1), \qquad (39)$$

with probability at least $1-O(n^{-2})$. We now show (18) holds under the event defined in (39). Otherwise, there exists some $\tau_0\in J(\mathcal{P}_0)$ such that $|\hat{\tau}-\tau_0|\ge\delta_{\min}$, for all $\hat{\tau}\in J(\widehat{\mathcal{P}})$. Under the event defined in (39), we obtain that

$$\int_{\tau_0-\delta_{\min}}^{\tau_0+\delta_{\min}}\|\theta_0(a)-\theta_{0,[\tau_0-\delta_{\min},\tau_0+\delta_{\min})}\|_2^2da=o(1). \qquad (40)$$

On the other hand, since $\theta_0(a)$ is a constant function on $[\tau_0-\delta_{\min},\tau_0)$ or $[\tau_0,\tau_0+\delta_{\min})$, we have

$$\begin{aligned}&\int_{\tau_0-\delta_{\min}}^{\tau_0+\delta_{\min}}\|\theta_0(a)-\theta_{0,[\tau_0-\delta_{\min},\tau_0+\delta_{\min})}\|_2^2da\\ \geq\;& \min_{\theta\in\mathbb{R}^{p+1}}\left(\delta_{\min}\|\theta_{0,[\tau_0-\delta_{\min},\tau_0)}-\theta\|_2^2+\delta_{\min}\|\theta_{0,[\tau_0,\tau_0+\delta_{\min})}-\theta\|_2^2\right)\\ \geq\;& \frac{\delta_{\min}}{2}\|\theta_{0,[\tau_0-\delta_{\min},\tau_0)}-\theta_{0,[\tau_0,\tau_0+\delta_{\min})}\|_2^2\geq\frac{\delta_{\min}\kappa_0^2}{2},\end{aligned}$$

where

$$\kappa_0\equiv\min_{\substack{\mathcal{I}_1,\mathcal{I}_2\in\mathcal{P}_0\\ \mathcal{I}_1\text{ and }\mathcal{I}_2\text{ are adjacent}}}\|\theta_{0,\mathcal{I}_1}-\theta_{0,\mathcal{I}_2}\|_2>0.$$

This apparently violates (40). (18) thus holds with probability at least $1-O(n^{-2})$.
Part 2: By (32) and (33), we have with probability at least $1-O(n^{-2})$ that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^n\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\widehat{\mathcal{P}}|\geq\eta_1+n\gamma_n|\widehat{\mathcal{P}}|-2c_0^2|\widehat{\mathcal{P}}|\log n.$$

Notice that

$$\begin{aligned}\eta_1\;&=\;\eta_4+2\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^n\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\{\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta_{0,\mathcal{I}}\}\\ &+\;\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^n\mathbb{I}(A_i\in\mathcal{I})\{\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta_{0,\mathcal{I}}\}^2.\end{aligned}$$

Denoted by $\mathfrak{T}(m)$ the set of intervals $\mathcal{I}\in\mathfrak{I}(m)$ with $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\geq\bar{c}_1n^{-1}\log n$. Under the events defined in Lemma 2 and 3, we have

$$\begin{aligned}\eta_1\;&\geq\;\eta_4+2\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^n\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\{\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta_{0,\mathcal{I}}\}\\ &+\;\sum_{\mathcal{I}\in\widehat{\mathcal{P}},\mathcal{I}\in\mathfrak{T}(m)}\sum_{i=1}^n\mathbb{I}(A_i\in\mathcal{I})\{\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta_{0,\mathcal{I}}\}^2\geq\eta_4-2\bar{c}_2|\widehat{\mathcal{P}}|\log n\\ &+\;\sum_{\mathcal{I}\in\widehat{\mathcal{P}},\mathcal{I}\in\mathfrak{T}(m)}\left(\frac{n}{c_1}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da-3c_1\sqrt{n\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\log n}\right).\end{aligned}$$

To summarize, we've shown that with probability at least $1-O(n^{-2})$,

$$\begin{aligned}&\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^n\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\widehat{\mathcal{P}}| \qquad (41)\\ \geq\;&\sum_{\mathcal{I}\in\widehat{\mathcal{P}},\mathcal{I}\in\mathfrak{T}(m)}\left(\frac{n}{c_1}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da-3c_1\sqrt{n\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\log n}\right)\\ +\;&\eta_4+n\gamma_n|\widehat{\mathcal{P}}|-2(c_0^2+\bar{c}_2)|\widehat{\mathcal{P}}|\log n.\end{aligned}$$

It follows from (37) and (38) that

$$\begin{aligned}&\eta_4 + n\lambda_n \sup_{\mathcal{I}\in\mathcal{P}_0} \|\theta_{0,\mathcal{I}}\|_2^2 + n\gamma_n|\mathcal{P}_0|\\ \geq\ & \left\{\sum_{\mathcal{I}\in\mathcal{P}^*}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{\mathcal{I}}^*)^2+n\lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}^*\|_2^2\right)+n\gamma_n|\mathcal{P}^*|\right\}-c_0^*\log n,\end{aligned}$$

for some constants $c_0^*>0$, with probability at least $1-O(n^{-2})$. By (27) and the condition that $\lambda_n=O(n^{-1}\log n)$, there exists some constant $c_1^*>c_0^*$ such that

$$\begin{aligned}&\eta_4 + n\gamma_n|\mathcal{P}_0|\\ \geq\ & \left\{\sum_{\mathcal{I}\in\mathcal{P}^*}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{\mathcal{I}}^*)^2+n\lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}^*\|_2^2\right)+n\gamma_n|\mathcal{P}^*|\right\}-c_1^*\log n,\end{aligned}$$

with probability at least $1-O(n^{-2})$. In view of (41), we've shown that with probability at least $1-O(n^{-2})$,

$$\begin{aligned}&\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\widehat{\mathcal{P}}| \qquad (42)\\ \geq\ & \sum_{\mathcal{I}\in\widehat{\mathcal{P}},\mathcal{I}\in\mathfrak{I}(m)}\left(\frac{n}{c_1}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da-3c_1\sqrt{n\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\log n}\right)\\ +\ & \left\{\sum_{\mathcal{I}\in\mathcal{P}^*}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{\mathcal{I}}^*)^2+n\lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}^*\|_2^2\right)+n\gamma_n|\mathcal{P}^*|\right\}\\ +\ & n\gamma_n|\widehat{\mathcal{P}}|-(2c_0^2+2\bar{c}_2)|\widehat{\mathcal{P}}|\log n-c_1^*\log n-n\gamma_n|\mathcal{P}_0|.\end{aligned}$$

By definition,

$$\begin{aligned}&\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\widehat{\mathcal{P}}|\\ \leq\ & \sum_{\mathcal{I}\in\mathcal{P}^*}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{\mathcal{I}}^*)^2+n\lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}^*\|_2^2\right)+n\gamma_n|\mathcal{P}^*|.\end{aligned}$$

Thus, we have with probability at least $1-O(n^{-2})$,

$$\begin{aligned}&\sum_{\mathcal{I}\in\widehat{\mathcal{P}},\mathcal{I}\in\mathfrak{I}(m)}\left(\frac{n}{c_1}\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da-3c_1\sqrt{n\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\log n}\right)\\ &\qquad\leq (2c_0^2+2\bar{c}_2)|\widehat{\mathcal{P}}|\log n+c_1^*\log n+n\gamma_n|\mathcal{P}_0|-n\gamma_n|\widehat{\mathcal{P}}|,\end{aligned}$$

and hence,

$$\begin{aligned}&\sum_{\mathcal{I}\in\widehat{\mathcal{P}},\mathcal{I}\in\mathfrak{I}(m)}\frac{n}{c_1}\left(\sqrt{\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da}-\frac{3c_1}{2}n^{-1/2}\sqrt{\log n}\right)^2 \qquad (43)\\ &\leq (2c_0^2+2\bar{c}_2+9c_1/4)|\widehat{\mathcal{P}}|\log n+c_1^*\log n+n\gamma_n|\mathcal{P}_0|-n\gamma_n|\widehat{\mathcal{P}}|.\end{aligned}$$

Under the event defined in (19), we have either $|\widehat{\mathcal{P}}| \ge 2|\mathcal{P}_0|$, or $|\mathcal{P}_0| \le |\widehat{\mathcal{P}}| \le 2|\mathcal{P}_0|$. When $|\widehat{\mathcal{P}}| \ge 2|\mathcal{P}_0|$, it follows from the condition $n\gamma_n \gg \log n$ that for sufficiently large n, $\gamma_n/4 \ge 2c_0^2 + 2\bar{c}_2 + 9c_1/4$, $n|\mathcal{P}_0|\gamma_n \ge 2c_1^* \log n$ and hence

$$\begin{aligned} & (2c_0^2 + 2\bar{c}_2 + 9c_1/4)|\widehat{\mathcal{P}}| \log n + c_1^* \log n + n\gamma_n|\mathcal{P}_0| - n\gamma_n|\widehat{\mathcal{P}}| \\ \le\ & (2c_0^2 + 2\bar{c}_2 + 9c_1/4)|\widehat{\mathcal{P}}| \log n + c_1^* \log n - n\gamma_n|\widehat{\mathcal{P}}|/4 - n\gamma_n|\mathcal{P}_0|/2 \\ \le\ & (2c_0^2 + 2\bar{c}_2 + 9c_1/4)|\widehat{\mathcal{P}}| \log n - n\gamma_n|\widehat{\mathcal{P}}|/4 \le 0, \end{aligned}$$

When $|\mathcal{P}_0| \le |\widehat{\mathcal{P}}| \le 2|\mathcal{P}_0|$, we have

$$\begin{aligned} & (2c_0^2 + 2\bar{c}_2 + 9c_1/4)|\widehat{\mathcal{P}}| \log n + c_1^* \log n + n\gamma_n|\mathcal{P}_0| - n\gamma_n|\widehat{\mathcal{P}}| \\ \le\ & 2(2c_0^2 + 2\bar{c}_2 + 9c_1/4)|\mathcal{P}_0| \log n + c_1^* \log n. \end{aligned}$$

In view of (43), we have with probability at least $1 - O(n^{-2})$,

$$\sum_{\mathcal{I} \in \widehat{\mathcal{P}}, \mathcal{I} \in \mathfrak{I}(m)} \frac{n}{c_1} \left(\sqrt{\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da} - \frac{3c_1}{2} n^{-1/2} \sqrt{\log n} \right)^2 \le c \log n,$$

for some constant $c > 0$. Thus, with probability at least $1 - O(n^{-2})$, we have

$$\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da = O(n^{-1} \log n),$$

for any $\mathcal{I} \in \widehat{\mathcal{P}} \cap \mathfrak{I}(m)$. By the definition of $\mathfrak{I}(m)$, we obtain that with probability at least $1 - O(n^{-2})$,

$$\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da = O(n^{-1} \log n), \quad \forall \mathcal{I} \in \widehat{\mathcal{P}}. \tag{44}$$

Consider a given change point $\tau \in \mathcal{P}_0$, there exists an interval $\mathcal{I} \in \widehat{\mathcal{P}}$ of the form $[i_1, i_2)$ or $[i_1, i_2]$ with $i_2 = 1$ such that $i_1 \le \tau < i_2$. Under the event defined in (44), it follows from Lemma 3 such that $\min(|i_1 - \tau|, |i_2 - \tau|) = O(n^{-1} \log n)$. This proves (20).
Part 3: Using similar arguments in proving (30), we can show that the following events occur with probability at least $1 - O(n^{-2})$: for any interval $\mathcal{I} \in \widehat{\mathcal{P}}$, we have

$$\left| \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I}) \{Y_i - \overline{X}_i^\top \theta_0(A_i)\} \overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\} \right| \le C \log n,$$

for some constant $C > 0$.

By (44), using similar arguments in proving (41) and (42), we can show the following event occurs with probability at least $1 - O(n^{-2})$,

$$\begin{aligned} & \sum_{\mathcal{I} \in \widehat{\mathcal{P}}} \left(\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2 \right) + n\gamma_n|\widehat{\mathcal{P}}| \\ \ge\ & \left\{ \sum_{\mathcal{I} \in \mathcal{P}^*} \left(\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{\mathcal{I}}^*)^2 + n\lambda_n|\mathcal{I}|\|\theta_{\mathcal{I}}^*\|_2^2 \right) + n\gamma_n|\mathcal{P}^*| \right\} \\ +\ & n\gamma_n|\widehat{\mathcal{P}}| - C|\widehat{\mathcal{P}}| \log n - C \log n - n\gamma_n|\mathcal{P}_0|, \end{aligned}$$

for some constant $C > 0$. By definition,

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left(\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\widehat{\mathcal{P}}|$$

$$\leq \sum_{\mathcal{I}\in\mathcal{P}^*}\left(\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta^*_{\mathcal{I}})^2+n\lambda_n|\mathcal{I}|\|\theta^*_{\mathcal{I}}\|_2^2\right)+n\gamma_n|\mathcal{P}^*|.$$

Thus, we have with probability at least $1-O(n^{-2})$,

$$n\gamma_n|\widehat{\mathcal{P}}|-C|\widehat{\mathcal{P}}|\log n-C\log n-n\gamma_n|\mathcal{P}_0|\leq 0.$$

Since $\gamma_n \gg n^{-1}\log n$, the above event occurs only when $|\widehat{\mathcal{P}}|\leq|\mathcal{P}_0|$. To see this, notice that if $|\widehat{\mathcal{P}}|>|\mathcal{P}_0|$, we have

$$\begin{aligned}n\gamma_n-C\log n-\frac{C\log n}{|\widehat{\mathcal{P}}|}-n\gamma_n\frac{|\mathcal{P}_0|}{|\widehat{\mathcal{P}}|}&\geq n\gamma_n-C\log n-\frac{C\log n}{|\mathcal{P}_0|+1}-n\gamma_n\frac{|\mathcal{P}_0|}{|\mathcal{P}_0|+1}\\&=\frac{n\gamma_n}{|\mathcal{P}_0|+1}-C\log n-\frac{C\log n}{|\mathcal{P}_0|+1}\gg 0,\end{aligned}$$

since $\gamma_n \gg n^{-1}\log n$. This proves (21).
Part 4: In the first three parts, we've shown that

$$|\widehat{\mathcal{P}}|=|\mathcal{P}_0| \quad\text{and}\quad \max_{\tau\in J(\mathcal{P}_0)}\min_{\hat{\tau}\in J(\widehat{\mathcal{P}})}|\hat{\tau}-\tau|=O(n^{-1}\log n), \tag{45}$$

with probability tending to 1. For sufficiently large n, the event defined in (45) implies that $|\mathcal{I}|\geq\bar{c}_0n^{-1}\log n$ for any $\mathcal{I}\in\widehat{\mathcal{P}}$. Thus, it follows from Lemma 1 that the following occurs with probability at least $1-O(n^{-2})$: for any $\mathcal{I}\in\widehat{\mathcal{P}}$, we have

$$|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2^2\leq c_0^2n^{-1}\log n. \tag{46}$$

Under the events defined in (44), (45) and (46), we have

$$\int_0^1\|\widehat{\theta}(a)-\theta_0(a)\|_2^2da=\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\int_{\mathcal{I}}\|\widehat{\theta}_{\mathcal{I}}-\theta_0(a)\|_2^2da=\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\int_{\mathcal{I}}\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}+\theta_{0,\mathcal{I}}-\theta_0(a)\|_2^2da$$

$$=\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\int_{\mathcal{I}}\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2^2da+\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\int_{\mathcal{I}}\|\theta_{0,\mathcal{I}}-\theta_0(a)\|_2^2da+2\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\int_{\mathcal{I}}(\widehat{\theta}_{0,\mathcal{I}}-\theta_{0,\mathcal{I}})^\top\{\theta_{0,\mathcal{I}}-\theta_0(a)\}da$$

$$\leq 2\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\int_{\mathcal{I}}\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2^2da+2\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\int_{\mathcal{I}}\|\theta_{0,\mathcal{I}}-\theta_0(a)\|_2^2da=O(|\widehat{\mathcal{P}}|n^{-1}\log n)=O(|\mathcal{P}_0|n^{-1}\log n),$$

where the first inequality is due to Cauchy-Schwarz inequality. This proves (iii). The proof is hence completed.

B.2 Proof of Lemma 1

Proof of (22)*:* By definition, we have

$$
\begin{aligned}
&\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2 \\
\le\ & \left\|\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top+\lambda_n|\mathcal{I}|\mathbb{E}_{p+1}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_iY_i-\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}Y\right)\right\|_2 \\
+\ & \left\|\left\{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top+\lambda_n|\mathcal{I}|\mathbb{E}_{p+1}\right)^{-1}-\left(\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\overline{X}^\top\right)^{-1}\right\}\{\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}Y\}\right\|_2 \\
\le\ & \underbrace{\left\|\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top+\lambda_n|\mathcal{I}|\mathbb{E}_{p+1}\right)^{-1}\right\|_2}_{\eta_1(\mathcal{I})}\underbrace{\left\|\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_iY_i-\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}Y\right)\right\|_2}_{\eta_2(\mathcal{I})} \\
+\ & \underbrace{\left\|\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top+\lambda_n|\mathcal{I}|\mathbb{E}_{p+1}\right)^{-1}-\left(\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\overline{X}^\top\right)^{-1}\right\|_2}_{\eta_3(\mathcal{I})}\underbrace{\|\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}Y\|_2}_{\eta_4(\mathcal{I})}.
\end{aligned}
$$

It follows from Cauchy-Schwarz inequality that

$$\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2^2\le 2\eta_1^2(I)\eta_2^2(I)+2\eta_3^2(I)\eta_4^2(I). \tag{47}$$

In the following, we provides upper bounds for

$$\max_{\substack{\mathcal{I}\in\mathfrak{I}(m)\\ |\mathcal{I}|\ge\bar{c}_0n^{-1}\log n}}\eta_j(\mathcal{I}),$$

for $j=1,2,3,4$, where the constant $\bar{c}_0$ will be specified later. The uniform convergence rates of $\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2$ can thus be derived.

Without loss of generality, assume the constant ω in Condition (A4) is greater than or equal to $\log^{-1/2}2$. Then, we have $\exp(1/\omega^2)\le\exp(\log 2)=2$ and hence $\|1\|_{\psi_2|A}\le\omega$. Therefore, we have $\max_{j\in\{1,\dots,p+1\}}\|\overline{X}^{(j)}\|_{\psi_2|A}\le\omega$, almost surely. By the definition of the conditional Orlicz norm, this implies that

$$\mathrm{E}\left\{1+\sum_{q=1}^{+\infty}\frac{|\overline{X}^{(j)}|^{2q}}{\omega^{2q}q!}\middle|A\right\}\le 2,\qquad\forall j\in\{1,\dots,p+1\},$$

almost surely, and hence

$$\mathrm{E}(|\overline{X}^{(j)}|^{2q}|A)\le q!\omega^{2q},\qquad\forall j\in\{1,\dots,p+1\},q=1,2,\dots \tag{48}$$

By Cauchy-Schwarz inequality, we obtain that

$$\mathrm{E}(|\overline{X}^{(j_1)}\overline{X}^{(j_2)}|^q|A)\le\sqrt{\mathrm{E}(|\overline{X}^{(j_1)}|^{2q}|A)\mathrm{E}(|\overline{X}^{(j_2)}|^{2q}|A)}\le q!\omega^{2q}, \tag{49}$$

for any $j_1, j_2 \in \{1, \dots, p+1\}$ and any integer $q \geq 1$, almost surely.

Since A has a bounded probability density function $p_A(\cdot)$ in $[0,1]$, there exists some constant $C_0 > 0$ such that

$$\sup_{a\in[0,1]} p_A(a) \leq C_0 \text{ and } \Pr(A \in \mathcal{I}) \leq C_0|\mathcal{I}|, \tag{50}$$

for any interval $\mathcal{I} \in [0,1]$. This together with (49) yields that for any integer $q \geq 1$, $j_1, j_2 \in \{1, \dots, p+1\}$ and any interval $\mathcal{I} \in [0,1]$, we have

$$\mathrm{E}|\mathbb{I}(A \in \mathcal{I})\overline{X}^{(j_1)}\overline{X}^{(j_2)}|^q = \mathrm{E}\{\mathbb{I}(A \in \mathcal{I})\mathrm{E}(|\overline{X}^{(j_1)}\overline{X}^{(j_2)}|^q|A)\} \leq q!\omega^{2q}\mathrm{E}\mathbb{I}(A \in \mathcal{I}) \leq C_0 q!\omega^{2q}|\mathcal{I}|.$$

It follows that

$$\begin{aligned} & \mathrm{E}|\mathbb{I}(A \in \mathcal{I})\overline{X}^{(j_1)}\overline{X}^{(j_2)} - \mathrm{E}\mathbb{I}(A \in \mathcal{I})\overline{X}^{(j_1)}\overline{X}^{(j_2)}|^q \\ \leq\; & \mathrm{E}|\mathbb{I}(A_1 \in \mathcal{I})\overline{X}_1^{(j_1)}\overline{X}_1^{(j_2)} - \mathbb{I}(A_2 \in \mathcal{I})\overline{X}_2^{(j_1)}\overline{X}_2^{(j_2)}|^q \\ \leq\; & 2^{q-1}\mathrm{E}|\mathbb{I}(A_1 \in \mathcal{I})\overline{X}_1^{(j_1)}\overline{X}_1^{(j_2)}|^q + 2^{q-1}\mathrm{E}|\mathbb{I}(A_2 \in \mathcal{I})\overline{X}_2^{(j_1)}\overline{X}_2^{(j_2)}|^q \\ =\; & 2^q\mathrm{E}|\mathbb{I}(A \in \mathcal{I})\overline{X}^{(j_1)}\overline{X}^{(j_2)}|^q \leq C_0 q!(2\omega^2)^q|\mathcal{I}|, \end{aligned}$$

where the second inequality follows from Jensen's inequality and the third inequality is due to that $|a+b|^q \leq 2^{q-1}|a|^q + 2^{q-1}|b|^{q-1}$, for any $a, b \in \mathbb{R}$ and $q \geq 1$.

By the Bernstein's inequality (see Lemma 2.2.11, van der Vaart and Wellner, 1996), we obtain that

$$\begin{aligned} & \Pr\left(\left|\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\overline{X}_i^{(j_1)}\overline{X}_i^{(j_2)} - n\mathrm{E}\mathbb{I}(A \in \mathcal{I})\overline{X}^{(j_1)}\overline{X}^{(j_2)}\right| \geq t\omega^2\sqrt{|\mathcal{I}|n\log n}\right) \qquad (51) \\ \leq\; & 2\exp\left(-\frac{1}{2}\frac{\omega^4 t^2|\mathcal{I}|n\log n}{8nC_0\omega^4|\mathcal{I}| + 2\omega^4 t\sqrt{|\mathcal{I}|n\log n}}\right) \leq 2\exp\left(-\frac{t^2\log n}{16C_0 + 4t(n|\mathcal{I}|)^{-1/2}\sqrt{\log n}}\right), \end{aligned}$$

for any $t > 0$, any integers $j_1, j_2 \in \{1, \dots, p+1\}$ and any interval $\mathcal{I} \in [0,1]$. Set $t = 20\sqrt{C_0}$, for any interval $\mathcal{I}$ with $|\mathcal{I}| \geq C_0^{-1}n^{-1}\log n$, we have

$$t^2\log n \geq 4\{16C_0 + 4t(n|\mathcal{I}|)^{-1/2}\sqrt{\log n}\}.$$

It follows from (51) that

$$\Pr\left(\left|\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\overline{X}_i^{(j_1)}\overline{X}_i^{(j_2)} - n\mathrm{E}\mathbb{I}(A \in \mathcal{I})\overline{X}^{(j_1)}\overline{X}^{(j_2)}\right| \geq 20\omega^2\sqrt{C_0|\mathcal{I}|n\log n}\right) \leq 2n^{-4},$$

for any integers $j_1, j_2 \in \{1, \dots, p+1\}$ and any interval $\mathcal{I}$ that satisfies $|\mathcal{I}| \geq C_0^{-1}n^{-1}\log n$. Notice that the number of elements in $\mathfrak{I}(m)$ is bounded by $(m+1)^2$. Since $m \asymp n$, it follows from Bonferroni's inequality that

$$\Pr(\mathcal{A}_1) \geq 1 - 2(m+1)^2(p+1)^2n^{-4} = 1 - O(n^{-2}), \tag{52}$$

where the event $\mathcal{A}_1$ is defined as

$$\bigcap_{\substack{j_1,j_2\in\{1,\dots,p+1\}\\ \mathcal{I}\in\mathfrak{I}(m)\\ |\mathcal{I}|\ge C_0^{-1}n^{-1}\log n}} \left\{\left|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i^{(j_1)}\overline{X}_i^{(j_2)} - n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}^{(j_1)}\overline{X}^{(j_2)}\right| \le 20\omega^2\sqrt{C_0|\mathcal{I}|n\log n}\right\}.$$

For any symmetric matrix $\boldsymbol{A}$, we have $\|\boldsymbol{A}\|_2 \le \sqrt{\|\boldsymbol{A}\|_\infty\|\boldsymbol{A}\|_1} = \|\boldsymbol{A}\|_\infty$. Thus, under the event defined in $\mathcal{A}_1$, we have

$$\begin{aligned}\left\|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top - n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\overline{X}^\top\right\|_2 &\le \left\|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top - n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\overline{X}^\top\right\|_\infty \\ &\le 20\omega^2(p+1)\sqrt{C_0|\mathcal{I}|n\log n},\end{aligned}$$

for any $\mathcal{I}\in\mathfrak{I}(m)$ with $|\mathcal{I}|\ge C_0^{-1}n^{-1}\log n$. Since $\lambda_n = O(n^{-1}\log n)$, we obtain

$$\begin{aligned}&\left\|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top - n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\overline{X}^\top + n\lambda_n\mathbb{E}|\mathcal{I}|\right\|_2 \\ \le\ & n\lambda_n|\mathcal{I}| + \left\|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top - n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\overline{X}^\top\right\|_2 \le c\sqrt{|\mathcal{I}|n\log n},\end{aligned}$$

for some constant $c>0$. To summarize, under the event defined in $\mathcal{A}_1$, we've shown that

$$\left\|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top - n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\overline{X}^\top + n\lambda_n\mathbb{E}|\mathcal{I}|\right\|_2 \le c\sqrt{|\mathcal{I}|n\log n}, \tag{53}$$

for any interval $\mathcal{I}\in\mathfrak{I}(m)$ with $|\mathcal{I}|\ge C_0^{-1}n^{-1}\log n$.

Let $\Sigma = \mathrm{E}\overline{X}\overline{X}^\top$. If Σ is singular, there exists some nonzero vector $a\in\mathbb{R}^p$ and some $b\in\mathbb{R}$ such that $a^\top X = b$, almost surely. As a result, the covariance matrix of X is degenerate. Thus, we've reached a contraction. Therefore, Σ is nonsingular. There exists some constant $\bar{c}_*>0$ such that

$$\lambda_{\min}(\Sigma)\ge \bar{c}_*. \tag{54}$$

By (A3), we have

$$\Pr(A\in\mathcal{I}|X)\ge c_*|\mathcal{I}|,$$

for any interval $\mathcal{I}\in[0,1]$. This together with (54) implies that

$$\begin{aligned}\lambda_{\min}\left(\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\overline{X}^\top\right) &= \lambda_{\min}\left(\mathrm{E}\Pr(A\in\mathcal{I}|X)\overline{X}\overline{X}^\top\right) \\ &\ge c_*\lambda_{\min}(\mathrm{E}\overline{X}\overline{X}^\top)|\mathcal{I}| \ge c_*\bar{c}_*|\mathcal{I}|.\end{aligned} \tag{55}$$

For any interval $\mathcal{I}$ with $|\mathcal{I}|\ge 4c^2(c_*\bar{c}_*)^{-2}n^{-1}\log n$, we have

$$c_*\bar{c}_*|\mathcal{I}| - c\sqrt{|\mathcal{I}|n^{-1}\log n} \ge \frac{c_*\bar{c}_*|\mathcal{I}|}{2}.$$

In view of (53) and (55), we obtain that

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top+\lambda_n\mathbb{E}|\mathcal{I}|\right)\geq\lambda_{\min}\left(\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\,\overline{X}^\top\right) \tag{56}$$
$$-\quad\frac{1}{n}\left\|\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top-n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\,\overline{X}^\top+n\lambda_n\mathbb{E}|\mathcal{I}|\right\|_2\geq\frac{c_*\bar{c}_*|\mathcal{I}|}{2}.$$

Set $\bar{c}_0=\max(4c^2(c_*\bar{c}_*)^{-1},C_0^{-1})$, it is immediate to see that

$$\max_{\substack{\mathcal{I}\in\mathfrak{I}(m)\\|\mathcal{I}|\geq\bar{c}_0n^{-1}\log n}}\eta_1(\mathcal{I})\leq\frac{2}{c_*\bar{c}_*|\mathcal{I}|}, \tag{57}$$

under the event defined in $\mathcal{A}_1$.

For any $\mathcal{I}\in[0,1]$, we have

$$\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top+\lambda_n\mathbb{E}|\mathcal{I}|\right)^{-1}-\left(\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\,\overline{X}^\top\right)^{-1}\right\|_2$$
$$\leq\quad\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top+\lambda_n\mathbb{E}|\mathcal{I}|\right)^{-1}\right\|_2\left\|\left(\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\,\overline{X}^\top\right)^{-1}\right\|_2$$
$$\times\quad\left\|\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top-n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}\,\overline{X}^\top+n\lambda_n\mathbb{E}|\mathcal{I}|\right\|_2$$

This together with (53), (55) and (56) yields

$$\max_{\substack{\mathcal{I}\in\mathfrak{I}(m)\\|\mathcal{I}|\geq\bar{c}_0n^{-1}\log n}}\eta_3(\mathcal{I})\leq\frac{2c\sqrt{n^{-1}\log n}}{c_*^2\bar{c}_*^2|\mathcal{I}|^{3/2}}, \tag{58}$$

under the event defined in $\mathcal{A}_1$.

Similar to (49), we can show that for any integer $q\geq1$ and $j\in\{1,\dots,p+1\}$,

$$\mathrm{E}(|\overline{X}^{(j)}Y|^q|A)\leq q!\omega^{2q}, \tag{59}$$

almost surely. Specifically, set $q=1$, we obtain $\mathrm{E}(|\overline{X}^{(j)}Y||A)\leq\omega^2$. By (50), we have that

$$\|\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}Y\|_2\leq\left(\sum_{j=1}^{p+1}|\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}^{(j)}Y|^2\right)^{1/2}\leq\left(\sum_{j=1}^{p+1}|\mathrm{E}\{\mathbb{I}(A\in\mathcal{I})\mathrm{E}(|\overline{X}^{(j)}Y||A)\}|^2\right)^{1/2}$$
$$\leq\left(\sum_{j=1}^{p+1}|\omega^2\mathrm{E}(A\in\mathcal{I})|^2\right)^{1/2}\leq C_0\sqrt{p+1}\omega^2|\mathcal{I}|.$$

for any $\mathcal{I} \in [0,1]$. This implies that

$$\max_{\substack{\mathcal{I}\in\mathfrak{I}(m)\\ |\mathcal{I}|\ge \bar{c}_0 n^{-1}\log n}} \eta_4(\mathcal{I}) \le C_0\sqrt{p+1}\omega^2|\mathcal{I}|. \tag{60}$$

Moreover, in view of (51) and (52), we can similarly show that

$$\Pr(\mathcal{A}_2) \ge 1-2(m+1)^2(p+1)n^{-4} = 1-O(n^{-2}), \tag{61}$$

where the event $\mathcal{A}_2$ is defined as

$$\bigcap_{\substack{j\in\{1,\dots,p+1\}\\ \mathcal{I}\in\mathfrak{I}(m)\\ |\mathcal{I}|\ge \bar{c}_0 n^{-1}\log n}} \left\{\left|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i^{(j)}Y_i - n\mathrm{E}\mathbb{I}(A\in\mathcal{I})\overline{X}^{(j)}Y\right| \le 20\omega^2\sqrt{C_0|\mathcal{I}|n\log n}\right\}.$$

Under the event defined in $\mathcal{A}_2$, we have

$$\max_{\substack{\mathcal{I}\in\mathfrak{I}(m)\\ |\mathcal{I}|\ge \bar{c}_0 n^{-1}\log n}} \eta_2(\mathcal{I}) \le 20\omega^2\sqrt{(p+1)C_0|\mathcal{I}|n^{-1}\log n}. \tag{62}$$

Combining (57) together with (58), (60), (62) yields

$$\max_{\substack{\mathcal{I}\in\mathfrak{I}(m)\\ |\mathcal{I}|\ge \bar{c}_0 n^{-1}\log n}} |\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2^2 = O\left(\frac{\log n}{n}\right),$$

under the events defined in $\mathcal{A}_1$ and $\mathcal{A}_2$. The proof is thus completed based on (52) and (61).

Proofs of (23), (24) *and* (27): We first prove (27). By the definition of $\theta_{0,\mathcal{I}}$, we have

$$\|\theta_{0,\mathcal{I}}\|_2 \le \left\|\left(\mathrm{E}\overline{X}\overline{X}^\top\mathbb{I}(A\in\mathcal{I})\right)^{-1}\right\|_2 \left\|\mathrm{E}\overline{X}Y\mathbb{I}(A\in\mathcal{I})\right\|_2.$$

It follows from (50), (55) and (59) that

$$\|\theta_{0,\mathcal{I}}\|_2 \le (c_*\bar{c}_*|\mathcal{I}|)^{-1}\sqrt{\sum_{j=1}^{p+1}\left[\mathrm{E}\left\{\left(\mathrm{E}|\overline{X}^{(j)}Y||A\right)\mathbb{I}(A\in\mathcal{I})\right\}\right]^2} \le \sqrt{p+1}(c_*\bar{c}_*)^{-1}C_0\omega^2,$$

for any $\mathcal{I}\in[0,1]$. Assertion (27) thus follows.

Consider (23). Since p is fixed, it suffices to show for any $j\in\{1,\dots,p+1\}$, the following event occurs with probability at least $1-O(n^{-2})$:

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta_{0,\mathcal{I}})\overline{X}_i^{(j)}\right|_2 = O\left(\frac{\sqrt{|\mathcal{I}|\log n}}{\sqrt{n}}\right). \tag{63}$$

By (27), (63) can be proven in a similar manner as (52) and (61). (24) can be similarly proven.

Proof of (25)*:* Similar to (23), we can show that the following event occurs with probability at least $1 - O(n^{-2})$: for any $\mathcal{I} \in \mathfrak{I}(m)$ such that $|\mathcal{I}| \ge \bar{c}_0 n^{-1} \log n$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})[\overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\}]^2 - \mathrm{E}\mathbb{I}(A \in \mathcal{I})[\overline{X}^\top \{\theta_0(A) - \theta_{0,\mathcal{I}}\}]^2 \right| = O\left(\frac{\sqrt{|\mathcal{I}| \log n}}{\sqrt{n}} \right).$$

Notice that

$$\begin{aligned} & \mathrm{E}\mathbb{I}(A \in \mathcal{I})[\overline{X}^\top \{\theta_0(A) - \theta_{0,\mathcal{I}}\}]^2 = \mathrm{E} \int_{\mathcal{I}} [\overline{X}^\top \{\theta_0(a) - \theta_{0,\mathcal{I}}\}]^2 \pi(a|X) da \\ \ge \;\; & c_* \mathrm{E} \int_{\mathcal{I}} [\overline{X}^\top \{\theta_0(a) - \theta_{0,\mathcal{I}}\}]^2 da = c_* \int_{\mathcal{I}} \{\theta_0(a) - \theta_{0,\mathcal{I}}\}^\top \Sigma \{\theta_0(a) - \theta_{0,\mathcal{I}}\} da \\ \ge \;\; & c_* \lambda_{\min}(\Sigma) \int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \ge c_* \bar{c}_* \int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da, \end{aligned} \tag{64}$$

where the first inequality is due to Condition (A3) and the last inequality is due to (54).

It follows that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})[\overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\}]^2 \ge c_* \bar{c}_* \int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da - O\left(\frac{\sqrt{|\mathcal{I}| \log n}}{\sqrt{n}} \right),$$

for any $\mathcal{I} \in \mathfrak{I}(m)$ such that $|\mathcal{I}| \ge \bar{c}_0 n^{-1} \log n$, with probability at least $1 - O(n^{-2})$. This completes the proof.

Proof of (26)*:* Similar to (23), we can show that the following event occurs with probability at least $1 - O(n^{-2})$: for any $\mathcal{I} \in \mathfrak{I}(m)$ such that $|\mathcal{I}| \ge \bar{c}_0 n^{-1} \log n$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})(|Y_i|^2 + \|\overline{X}_i\|_2^2) - \mathrm{E}\mathbb{I}(A \in \mathcal{I})(Y^2 + \|\overline{X}\|_2^2) \right| = O\left(\frac{\sqrt{|\mathcal{I}| \log n}}{\sqrt{n}} \right). \tag{65}$$

By (50) and (48), we have

$$\mathrm{E}\mathbb{I}(A \in \mathcal{I}) \|\overline{X}\|_2^2 \le \sum_{j=1}^{p+1} \mathrm{E}\mathbb{I}(A \in \mathcal{I}) |\overline{X}^{(j)}|^2 \le (p+1) C_0 \omega^2 |\mathcal{I}|.$$

Similarly, we can show

$$\mathrm{E}\mathbb{I}(A \in \mathcal{I}) Y^2 \le C_0 \omega^2 |\mathcal{I}|,$$

and thus

$$\mathrm{E}\mathbb{I}(A \in \mathcal{I})(Y^2 + \|\overline{X}\|_2^2) \le (p+2) C_0 \omega^2 |\mathcal{I}|.$$

This together with (65) yields (26).

B.3 Proof of Lemma 2

We first prove (28). By (27), we have $\sup_{a\in[0,1]}\|\theta_0(a)\|_2 \le c_0$ and hence

$$\sup_{a\in[0,1]}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2 \le 2c_0. \tag{66}$$

Similar to (48), we can show that for any integer $q\ge 1$,

$$\mathrm{E}(|Y|^{2q}|A)\le q!\omega^{2q}. \tag{67}$$

For any $\mathcal{I}\subseteq\mathfrak{I}(m)$ and integer $q\ge 2$, it follows from (59), (66) and (67) that

$$\begin{aligned}
&\mathrm{E}\left([Y\overline{X}^\top\{\theta_0(A)-\theta_{0,\mathcal{I}}\}]^q|A\right)\le\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^q\mathrm{E}(|Y|^q\|\overline{X}\|_2^q|A) \qquad\qquad (68)\\
\le\;& \frac{1}{2}\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^q\mathrm{E}\left(|Y|^{2q}+\left|\sum_{j=1}^{p+1}(\overline{X}^{(j)})^2\right|^q\middle|A\right)\le\frac{1}{2}\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^q q!\omega^{2q}\\
+\;& \frac{1}{2}\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^q(p+1)^{q-1}\sum_{j=1}^{p+1}\mathrm{E}(|\overline{X}^{(j)}|^{2q}|A)\le\frac{q!\omega^{2q}}{2}\{1+(p+1)^q\}\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^q\\
\le\;& q!\omega^{2q}(p+1)^q\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^q\le q!\omega^{2q}(p+1)^q(2c_0)^{q-2}\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^2.
\end{aligned}$$

Similarly, we can show

$$\mathrm{E}\left([\{\overline{X}^\top\theta_0(A)\}\overline{X}^\top\{\theta_0(A)-\theta_{0,\mathcal{I}}\}]^q|A\right)\le q!\omega^{2q}(p+1)^q2^{q-2}c_0^{2q-2}\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^2.$$

This together with (68) yields that for any integer $q\ge 2$, $\mathcal{I}\subseteq[0,1]$, we have

$$\mathrm{E}\left([\{Y-\overline{X}^\top\theta_0(A)\}\overline{X}^\top\{\theta_0(A)-\theta_{0,\mathcal{I}}\}]^q|A\right)\le q!c^q\|\theta_0(A)-\theta_{0,\mathcal{I}}\|_2^2, \tag{69}$$

for some constant $c>0$. Combining (50) together with (69), we obtain that for any integer $q\ge 2$, $\mathcal{I}\subseteq[0,1]$,

$$\begin{aligned}
\mathrm{E}[\mathbb{I}(A\in\mathcal{I})\{Y-\overline{X}^\top\theta_0(A)\}\overline{X}^\top\{\theta_0(A)-\theta_{0,\mathcal{I}}\}]^q &\le C_0q!c^q\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2p_A(a)da\\
&\le C_0q!c^q\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da.
\end{aligned}$$

Applying the Bernstein's inequality (using similar arguments in (51) and (52)), we can show that with probability at least $1-O(n^{-2})$, we have for any interval $\mathcal{I}$ that satisfies $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\ge(C_0)^{-1}n^{-1}\log n$ and $\mathcal{I}\in\mathfrak{I}(m)$,

$$\left|\sum_{i=1}^n\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i^\top\{\theta_0(A_i)-\theta_{0,\mathcal{I}}\}\right|\le O(1)\sqrt{n\log n}\left(\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\right)^{1/2},$$

where $O(1)$ denotes some positive constant. This proves (28).

Similarly, we can show that with probability at least $1-O(n^{-2})$, there exists some constant $C>0$ such that for any interval $\mathcal{I}$ that satisfies $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da \ge (C_0)^{-1}n^{-1}\log n$ and $\mathcal{I}\in\mathfrak{I}(m)$, we have

$$\left|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})[\overline{X}_i^\top\{\theta_0(A_i)-\theta_{0,\mathcal{I}}\}]^2 - n\mathrm{E}\mathbb{I}(A\in\mathcal{I})[\overline{X}^\top\{\theta_0(A)-\theta_{0,\mathcal{I}}\}]^2\right|$$
$$\le O(1)\sqrt{n\log n}\left(\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\right)^{1/2},$$

for some postive constant $O(1)$. This together with (64) yields (29).

B.4 Proof of Lemma 3

Consider the following three categories of intervals.
Category 1: Suppose i_1 and i_2 satisfy $\tau_{0,k-1}\le i_1\le i_2\le\tau_{0,k}$ for some integer k such that $1\le k\le K$. Then apparently, we have $\theta_{0,\mathcal{I}}=\theta_0(a)$, $\forall a\in\mathcal{I}$, and hence $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da = 0$. The assertion $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\le c_n$ is thus automatically satisfied.
Category 2: Suppose there exists some integer k such that $2\le k\le K$ and i_1,i_2 satisfy $\tau_{0,k-2}\le i_1<\tau_{0,k-1}<i_2\le\tau_{0,k}$. Assume we have

$$\min_{j\in\{1,2\}}|i_j-\tau_{0,k-1}|\ge\frac{3}{\kappa_0^2}c_n.$$

where

$$\kappa_0\equiv\min_{\substack{\mathcal{I}_1,\mathcal{I}_2\in\mathcal{P}_0\\ \mathcal{I}_1\text{ and }\mathcal{I}_2\text{ are adjacent}}}\|\theta_{0,\mathcal{I}_1}-\theta_{0,\mathcal{I}_2}\|_2>0.$$

Since $c_n\to 0$, for sufficiently large n, we have $\tau_{0,k}>\tau_{0,k-1}+3\kappa_0^{-2}c_n$ and $\tau_{0,k-2}+3\kappa_0^{-2}c_n<\tau_{0,k-1}$. Then, we have

$$\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\ge\min_{\theta\in\mathbb{R}^{p+1}}\int_{\tau_{0,k-1}-3\kappa_0^{-2}c_n}^{\tau_{0,k-1}+3\kappa_0^{-2}c_n}\|\theta-\theta_0(a)\|_2^2da$$
$$\ge\frac{6}{\kappa_0^{-2}}c_n\min_{\theta\in\mathbb{R}^{p+1}}\left(\|\theta-\theta_{0,[\tau_{0,k-2},\tau_{0,k-1})}\|_2^2,\|\theta-\theta_{0,[\tau_{0,k-1},\tau_{0,k})}\|_2^2\right)\ge\frac{6}{\kappa_0^{-2}}c_n\frac{\kappa_0^{-2}}{4}>c_n.$$

This violates the assertion that $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\le c_n$. We've thus reached a contradiction. As a result, we have

$$\min_{j\in\{1,2\}}|i_j-\tau_{0,k-1}|\le\frac{3}{\kappa_0^2}c_n.$$

Category 3: Suppose there exists some integer k such that $3\le k\le K$ and i_1,i_2 satisfy $\tau_{0,k-3}\le i_1<\tau_{0,k-2}<\tau_{0,k-1}<i_2\le\tau_{0,k}$. Assume we have

$$|i_1-\tau_{0,k-2}|\ge\frac{3}{\kappa_0^2}c_n.$$

Then for sufficiently large n, we have

$$\begin{aligned}
&\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \ge \min_{\theta \in \mathbb{R}^{p+1}} \int_{\tau_{0,k-2}-3\kappa_0^{-2}c_n}^{\tau_{0,k-2}+3\kappa_0^{-2}c_n} \|\theta - \theta_0(a)\|_2^2 da \\
\ge \;\; & \frac{6}{\kappa_0^{-2}} c_n \min_{\theta \in \mathbb{R}^{p+1}} \left(\|\theta - \theta_{0,[\tau_{0,k-2},\tau_{0,k-1})}\|_2^2, \|\theta - \theta_{0,[\tau_{0,k-1},\tau_{0,k})}\|_2^2 \right) \ge \frac{6}{\kappa_0^{-2}} c_n \frac{\kappa_0^{-2}}{4} > c_n.
\end{aligned}$$

This violates the assertion that $\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \le c_n$. We've thus reached a contradiction. As a result, we have $|i_1 - \tau_{0,k-2}| \le 3\kappa_0^{-2} c_n$. Similarly, we can show $|i_2 - \tau_{0,k-1}| \le 3\kappa_0^{-2} c_n$. Therefore, we obtain

$$\max_{j \in \{1,2\}} |i_j - \tau_{0,k-3+j}| \le \frac{3}{\kappa_0^2} c_n.$$

If $\mathcal{I}$ belongs to none of these categories, then there exists some integer k such that $2 \le k \le K$ and i_1, i_2 satisfy $i_1 \le \tau_{0,k-2}$ and $i_2 \ge \tau_{0,k}$. Using similar arguments, we can show that

$$\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \ge \int_{\tau_{0,k-2}}^{\tau_{0,k}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \ge \frac{\kappa_0^2}{4} \min_{\mathcal{I} \in \mathcal{P}_0} |\mathcal{I}|.$$

For sufficiently large n, this violates the assertion that $\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \le c_n$. We've thus reached a contradiction. Therefore, we shall have $\tau_{0,k-2} \le i_1 < i_2 \le \tau_{0,k}$. This completes the first part of the proof.

We now show (30). Take $c_n = \bar{c}_1 n^{-1} \log n$ and consider any interval $\mathcal{I} \in \mathfrak{I}(m)$ that satisfies $\int_{\mathcal{I}} \|\theta_0(a) - \theta_{0,\mathcal{I}}\|_2^2 da \le \bar{c}_1 n^{-1} \log n$.

If $\mathcal{I}$ belongs to Category 1, then $\theta_0(a) = \theta_{0,\mathcal{I}}$ for any $a \in \mathcal{I}$. As a result, we have

$$\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I}) \{Y_i - \overline{X}_i^\top \theta_0(A_i)\} \overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\} = 0.$$

If $\mathcal{I}$ belongs to Category 2, then there exists some integer k such that $2 \le k \le K$ and i_1, i_2 satisfy $\tau_{0,k-2} \le i_1 < \tau_{0,k-1} < i_2 \le \tau_{0,k}$. Thus, we have

$$\begin{aligned}
& \sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I}) \{Y_i - \overline{X}_i^\top \theta_0(A_i)\} \overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\} \\
= \;\; & \underbrace{\sum_{i=1}^{n} \mathbb{I}(A_i \in [i_1, \tau_{0,k-1})) \{Y_i - \overline{X}_i^\top \theta_0(A_i)\} \overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\}}_{\zeta_1} \\
+ \;\; & \underbrace{\sum_{i=1}^{n} \mathbb{I}(A_i \in [\tau_{0,k-1}, i_2)) \{Y_i - \overline{X}_i^\top \theta_0(A_i)\} \overline{X}_i^\top \{\theta_0(A_i) - \theta_{0,\mathcal{I}}\}}_{\zeta_2}.
\end{aligned}$$

Notice that we've shown

$$\min_{j \in \{1,2\}} |i_j - \tau_{0,k-1}| \le \frac{3\bar{c}_1}{\kappa_0^2} n^{-1} \log n.$$

Without loss of generality, suppose $|i_1-\tau_{0,k-1}|\le 3\bar{c}_1\kappa_0^{-2}n^{-1}\log n$. Using similar arguments in (37) and (38), we can show that $\zeta_1=O(\log n)$, with probability at least $1-O(n^{-2})$.

As for ζ_2, consider intervals of the form $[\tau_{0,j},(m+1)^{-1}i)$ for $j=0,1,\dots,K-1$, $i=1,\dots,m+1$. Denoted by $\mathfrak{J}(m)$ the set consisting of all such intervals. Similar to Lemma 1, we can show that the following event occurs with probability at least $1-O(n^{-2})$:

$$\left\|\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\right\|_2=O(\sqrt{n|\mathcal{I}|\log n}),\tag{70}$$

for any $\mathcal{I}\in\mathfrak{J}(m)$ with $|\mathcal{I}|\ge cn^{-1}\log n$ for some constant $c>0$. Suppose $i_2-\tau_{0,k-1}\ge cn^{-1}\log n$. Under the event defined in (70), it follows that

$$\left\|\sum_{i=1}^{n}\mathbb{I}(A_i\in[\tau_{0,k-1},i_2))\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\right\|_2=O(\sqrt{n|\mathcal{I}|\log n}),\tag{71}$$

Since $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\le\bar{c}_1n^{-1}\log n$, we have $\int_{\tau_{0,k-1}}^{i_2}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2da\le\bar{c}_1n^{-1}\log n$, and hence $(i_2-\tau_{0,k-1})\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2\le\bar{c}_1n^{-1}\log n$, for any $a\in[\tau_{0,k-1},i_2)$. This together with (71) yields that

$$\begin{aligned}&\left|\sum_{i=1}^{n}\mathbb{I}(A_i\in[\tau_{0,k-1},i_2))\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\{\theta_0(A_i)-\theta_{0,\mathcal{I}}\}\right|\\ \le\;&\left\|\sum_{i=1}^{n}\mathbb{I}(A_i\in[\tau_{0,k-1},i_2))\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\right\|_2\|\theta_0(\tau_{0,k-1})-\theta_{0,\mathcal{I}}\|_2=O(\log n),\end{aligned}$$

and hence $\zeta_2=O(\log n)$. When $i_2-\tau_{0,k-1}\le cn^{-1}\log n$, using similar arguments in (37) and (38), we can show that $\zeta_2=O(\log n)$, with probability at least $1-O(n^{-2})$. Thus, we've shown that with probability at least $1-O(n^{-2})$, for any interval $\mathcal{I}$ that belongs to the Category 2 with $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2\le\bar{c}_1n^{-1}\log n$, we have

$$\left|\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\{\theta_0(A_i)-\theta_{0,\mathcal{I}}\}\right|=O(\log n).$$

Similarly, one can show that with probability at least $1-O(n^{-2})$, for any interval $\mathcal{I}$ that belongs to the Category 3 with $\int_{\mathcal{I}}\|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2^2\le\bar{c}_1n^{-1}\log n$, we have

$$\left|\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\{\theta_0(A_i)-\theta_{0,\mathcal{I}}\}\right|=O(\log n).$$

The proof is thus completed.

B.5 Proof of Lemma 4

Consider a given interval $\mathcal{I}\in\widehat{\mathcal{P}}$. Suppose $|\mathcal{I}|<\bar{c}_3\gamma_n$. The value of the constant $\bar{c}_3$ will be determined later. Then, for sufficiently large n, we can find some interval $\mathcal{I}'\in\mathfrak{J}(m)\cap\widehat{\mathcal{P}}$

that is adjacent to $\mathcal{I}$. Thus, we have $\mathcal{I} \cup \mathcal{I}' \in \mathfrak{I}(m)$, and hence

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + \lambda_n |\mathcal{I}| \|\widehat{\theta}_{\mathcal{I}}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I}')(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2 \quad (72)$$
$$+ \quad \lambda_n |\mathcal{I}'| \|\widehat{\theta}_{\mathcal{I}'}\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}\cup\mathcal{I}'})^2 + \lambda_n |\mathcal{I} \cup \mathcal{I}'| \|\widehat{\theta}_{\mathcal{I}\cup\mathcal{I}'}\|_2^2 - \gamma_n.$$

Notice that the left-hand-side of the above expression is nonnegative. It follows that

$$\gamma_n \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}\cup\mathcal{I}'})^2 + \lambda_n |\mathcal{I} \cup \mathcal{I}'| \|\widehat{\theta}_{\mathcal{I}\cup\mathcal{I}'}\|_2^2.$$

By definition, we have

$$\widehat{\theta}_{\mathcal{I}\cup\mathcal{I}'} = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{p+1}} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')(Y_i - \overline{X}_i^\top \theta)^2 + \lambda_n |\mathcal{I} \cup \mathcal{I}'| \|\theta\|_2^2 \right). \quad (73)$$

Therefore, we obtain that

$$\begin{aligned} \gamma_n \quad & \leq \sum_{i=1}^{n} \frac{\mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')(Y_i - \overline{X}_i^\top \mathbf{0}_{p+1})^2}{n} + \lambda_n |\mathcal{I} \cup \mathcal{I}'| \|\mathbf{0}_{p+1}\|_2^2 \quad (74) \\ & = \sum_{i=1}^{n} \frac{\mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')Y_i^2}{n}. \end{aligned}$$

Suppose

$$|\mathcal{I} \cup \mathcal{I}'| \leq \frac{\gamma_n}{8c_0}, \quad (75)$$

where the constant c_0 is defined in Lemma 1.

Since $\gamma_n \gg n^{-1}$ and $m \asymp n$, we can find some interval $\mathcal{I}^* \in \mathfrak{I}(m)$ that covers $\mathcal{I} \cup \mathcal{I}'$ and satisfies $(8c_0)^{-1}\gamma_n \leq |\mathcal{I}^*| \leq (4c_0)^{-1}\gamma_n$. Under the event defined in (26), it follows from the condition $\gamma_n \gg n^{-1}\log n$ that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')Y_i^2 \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I}^*)Y_i^2 \leq c_0 \left(\frac{\sqrt{|(4c_0)^{-1}\gamma_n| \log n}}{\sqrt{n}} + (4c_0)^{-1}\gamma_n \right)$$
$$\leq 2c_0(4c_0)^{-1}\gamma_n = \frac{\gamma_n}{2},$$

for sufficiently large n. This apparently violates the results in (74). Thus, Assertion (75) doesn't hold. Therefore, we obtain that

$$|\mathcal{I} \cup \mathcal{I}'| \geq \frac{\gamma_n}{8c_0}, \quad (76)$$

with probability at least $1 - O(n^{-2})$.

Suppose the constant $\bar{c}_3$ satisfies $\bar{c}_3 \le (16c_0)^{-1}$. Under the event defined in (76), we have $|\mathcal{I}'| \ge \gamma_n (16c_0)^{-1}$. By (22), we have with probability at least $1-O(n^{-2})$ that $\|\widehat{\theta}_{\mathcal{I}'} - \theta_{0,\mathcal{I}'}\|_2 \le c_0\sqrt{n^{-1}\log n}|\mathcal{I}'|^{-1/2} \le 4c_0^{3/2}\sqrt{n^{-1}\log n}\gamma_n^{-1} \ll 1$. By (27), we have with probability at least $1-O(n^{-2})$ that

$$\|\widehat{\theta}_{\mathcal{I}'}\|_2 \le 2c_0, \tag{77}$$

for sufficiently large n.

In addition, it follows from (73) that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I}\cup\mathcal{I}')(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}\cup\mathcal{I}'})^2 + \lambda_n|\mathcal{I}\cup\mathcal{I}'|\|\widehat{\theta}_{\mathcal{I}\cup\mathcal{I}'}\|_2^2 \\ &\quad \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I}\cup\mathcal{I}')(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2 + \lambda_n|\mathcal{I}\cup\mathcal{I}'|\|\widehat{\theta}_{\mathcal{I}'}\|_2^2.\end{aligned}$$

By (72), this further implies that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + \lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2 \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2 + \lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}'}\|_2^2 - \gamma_n,$$

and hence

$$\gamma_n \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2 + \lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}'}\|_2^2.$$

By (77) and the conditions that $\lambda_n = O(n^{-1}\log n)$, $\gamma_n \gg n^{-1}\log n$, we have for sufficiently large n,

$$\frac{\gamma_n}{2} \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2.$$

It thus follows from Cauchy-Schwarz inequality and (77) that

$$\frac{\gamma_n}{2} \le \frac{2}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i^2 + \|\overline{X}_i^\top\|_2^2\|\widehat{\theta}_{\mathcal{I}'}\|_2^2) \le \frac{2(1+4c_0^2)}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i^2 + \|\overline{X}_i\|_2^2).$$

Using similar arguments in showing (76), we obtain that

$$|\mathcal{I}| \ge \frac{\gamma_n}{32(1+4c_0^2)c_0}.$$

with probability at least $1-O(n^{-2})$. Set $\bar{c}_3 = 32^{-1}(1+4c_0^2)^{-1}c_0^{-1}$, this violates the assumption that $|\mathcal{I}| < \bar{c}_3\gamma_n$. Thus, with probability at least $1-O(n^{-2})$, we obtain that $|\mathcal{I}| \ge \bar{c}_3\gamma_n$, for any $\mathcal{I} \in \widehat{\mathcal{P}}$. The proof is hence completed.

B.6 Proof of Theorem 2

Let $\{\widehat{\tau}_1, \widehat{\tau}_2, \dots, \widehat{\tau}_{\widehat{K}-1}\}$ be the set of change points in $J(\widehat{\mathcal{P}})$. Under the events defined in Theorem 1, we have $\widehat{K} = K$, and

$$\max_{k \in \{1,\dots,K-1\}} |\widehat{\tau}_k - \tau_{0,k}| \le c n^{-1} \log n, \tag{78}$$

for some constant $c > 0$. Set $\widehat{\tau}_0 = 0$ and $\widehat{\tau}_K = 1$.

Under the event defined in (78), we have for sufficiently large n that

$$\widehat{\tau}_k - \widehat{\tau}_{k-1} \ge \delta_{\min}, \quad \forall k \in \{1, \dots, K\}. \tag{79}$$

Since π^* satisfies $\sup_{\mathcal{I} \subseteq [0,1]} \sup_{a \in \mathcal{I}, x \in \mathbb{X}} |\mathcal{I}| \pi^*(a; x, \mathcal{I}) \asymp 1$, there exists some constant $\bar{c}_4 > 0$ such that $\pi^*(a; x, \widehat{d}(x)) \le \bar{c}_4 |\widehat{d}(x)|^{-1}$ for all a and x. This together with (79) yields that

$$\pi^*(a; x, \widehat{d}(x)) \le \bar{c}_4 \delta_{\min}^{-1}, \quad \forall a \in [0,1], x \in \mathbb{X}. \tag{80}$$

The rest of our proof is divided into three parts. In the first part, we show that there exists some constant $C > 0$ such that

$$\|\widehat{\theta}_{[\widehat{\tau}_{k-1}, \widehat{\tau}_k)} - \widehat{\theta}_{[\tau_{0,k-1}, \tau_{0,k})}\|_2 \le \frac{C \log n}{n}, \quad \forall k \in \{1, \dots, K\}, \tag{81}$$

with probability at least $1 - O(n^{-2})$. Using similar arguments in Lemma 1, we can show that there exists some constant $c_3 > 0$ such that the following events occur with probability at least $1 - O(n^{-2})$:

$$\|\widehat{\theta}_{[\tau_{0,k-1}, \tau_{0,k})} - \theta_{0,[\tau_{0,k-1}, \tau_{0,k})}\|_2 \le \frac{c_3 \sqrt{\log n}}{\sqrt{n \delta_{\min}}}, \quad \forall k \in \{1, \dots, K\}.$$

This together with (81) implies that

$$\|\widehat{\theta}_{[\widehat{\tau}_{k-1}, \widehat{\tau}_k)} - \theta_{0,[\tau_{0,k-1}, \tau_{0,k})}\|_2 \le \frac{2 c_3 \sqrt{\log n}}{\sqrt{n \delta_{\min}}}, \quad \forall k \in \{1, \dots, K\}, \tag{82}$$

for sufficiently large n, with probability at least $1 - O(n^{-2})$.

In the second part, we define an integer-valued function $\widehat{\mathbb{K}}(x)$ as follows. We set $\widehat{\mathbb{K}}(x) = k$ if $\widehat{d}(x) = [\widehat{\tau}_{k-1}, \widehat{\tau}_k)$ for some integer k such that $1 \le k \le K-1$, and set $\widehat{\mathbb{K}}(x) = K$ if $\widehat{d}(x) = [\widehat{\tau}_{K-1}, 1]$. By the definition of $\widehat{\theta}_{\mathcal{I}}$ and $\theta_{0,\mathcal{I}}$, we have almost surely $\widehat{\theta}_{[\widehat{\tau}_{K-1}, 1)} = \widehat{\theta}_{[\widehat{\tau}_{K-1}, 1]}$ and $\theta_{0,[\tau_{0,K-1}, 1)} = \theta_{0,[\tau_{0,K-1}, 1]}$. It is immediate to see that

$$\widehat{\mathbb{K}}(x) = \operatorname{sarg\,max}_{k \in \{1,\dots,K\}} \bar{x}^\top \widehat{\theta}_{[\widehat{\tau}_{k-1}, \widehat{\tau}_k)}, \tag{83}$$

where sarg max denotes the smallest maximizer when the argmax is not unique. In Part 2, we focus on proving

$$V^{\pi^*}(\widehat{d}) \ge \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(X)-1}, \tau_{0,\widehat{\mathbb{K}}(X)})}\right) - O(1) n^{-1} \log n, \tag{84}$$

with probability at least $1 - O(n^{-2})$, where $O(1)$ denotes some positive constant.

In the last part, we provide an opper bound for

$$V^{opt} - \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(X)-1},\tau_{0,\widehat{\mathbb{K}}(X)})}\right).$$

This together with (84) yields the desired results.

Proof of Part 1: Let $\widehat{\Delta}_k = [\widehat{\tau}_{k-1}, \widehat{\tau}_k) \cup [\tau_{0,k-1}, \tau_{0,k})^c + [\widehat{\tau}_{k-1}, \widehat{\tau}_k)^c \cup [\tau_{0,k-1}, \tau_{0,k})$. With some calculations, we can show that

$$\|\widehat{\theta}_{[\widehat{\tau}_{k-1},\widehat{\tau}_k)} - \widehat{\theta}_{[\tau_{0,k-1},\tau_{0,k})}\|_2 \le \zeta_1(k)\zeta_2(k) + \zeta_3(k)\zeta_4(k),$$

where

$$\begin{aligned}
\zeta_1(k) &= \left\| \left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}(\tau_{0,k-1} \le A_i < \tau_{0,k}) \overline{X}_i \overline{X}_i^\top + \lambda_n(\tau_{0,k} - \tau_{0,k-1}) \mathbb{E}_{p+1} \right)^{-1} \right\|_2, \\
\zeta_2(k) &= \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{I}(A_i \in \widehat{\Delta}_k) \overline{X}_i Y_i \right\|_2, \quad \zeta_3(k) = \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{I}(\tau_{0,k-1} \le A_i < \tau_{0,k}) \overline{X}_i Y_i \right\|_2, \\
\zeta_4(k) &= \left\| \left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}(\tau_{0,k-1} \le A_i < \tau_{0,k}) \overline{X}_i \overline{X}_i^\top + \lambda_n(\tau_{0,k} - \tau_{0,k-1}) \mathbb{E}_{p+1} \right)^{-1} \right. \\
&- \left. \left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}(\widehat{\tau}_{k-1} \le A_i < \widehat{\tau}_k) \overline{X}_i \overline{X}_i^\top + \lambda_n(\widehat{\tau}_k - \widehat{\tau}_{k-1}) \mathbb{E}_{p+1} \right)^{-1} \right\|_2.
\end{aligned}$$

Similar to (57), we can show with probability at least $1 - O(n^{-2})$ that

$$\max_{k \in \{1,\dots,K\}} \zeta_1(k) = O(1) \text{ and } \max_{k \in \{1,\dots,K\}} \zeta_5(k) = O(1), \tag{85}$$

where

$$\zeta_5(k) = \left\| \left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}(\widehat{\tau}_{k-1} \le A_i < \widehat{\tau}_k) \overline{X}_i \overline{X}_i^\top + \lambda_n(\widehat{\tau}_k - \widehat{\tau}_{k-1}) \mathbb{E}_{p+1} \right)^{-1} \right\|_2.$$

Under the event defined in (78), the Lebesgue measure of $\widehat{\Delta}_k$ is uniformly bounded by $2cn^{-1}\log n$, for any $k \in \{1,\dots,K\}$. Using similar arguments in (37) and (38), we can show with probability at least $1 - O(n^{-2})$ that

$$\max_{k \in \{1,\dots,K\}} \zeta_2(k) = O(n^{-1}\log n). \tag{86}$$

Similar to (60), we can show with probability at least $1 - O(n^{-2})$ that

$$\max_{k \in \{1,\dots,K\}} \zeta_3(k) = O(1). \tag{87}$$

Notice that $\zeta_4(k)$ can be upper bounded by

$$\begin{aligned}\zeta_4(k) &\le \zeta_1(k)\zeta_5(k)\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{I}(\tau_{0,k-1}\le A_i<\tau_{0,k})\overline{X}_i\overline{X}_i^\top+\lambda_n(\tau_{0,k}-\tau_{0,k-1})\mathbb{E}_{p+1}\right.\\ &- \left.\frac{1}{n}\sum_{i=1}^n \mathbb{I}(\widehat{\tau}_{k-1}\le A_i<\widehat{\tau}_{k})\overline{X}_i\overline{X}_i^\top-\lambda_n(\widehat{\tau}_{k}-\widehat{\tau}_{k-1})\mathbb{E}_{p+1}\right\|_2\\ &\le \zeta_1(k)\zeta_5(k)\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\widehat{\Delta}_k)\overline{X}_i\overline{X}_i^\top+\lambda_n(\tau_{0,k}-\tau_{0,k-1}-\widehat{\tau}_k+\widehat{\tau}_{k-1})\mathbb{E}_{p+1}\right\|_2.\end{aligned}$$

Under the condition $\lambda_n=O(n^{-1}\log n)$, using similar arguments in (37) and (38), we can show that with probability at least $1-O(n^{-2})$, the absolute value of each element in the matrix

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\widehat{\Delta}_k)\overline{X}_i\overline{X}_i^\top+\lambda_n(\tau_{0,k}-\tau_{0,k-1}-\widehat{\tau}_k+\widehat{\tau}_{k-1})\mathbb{E}_{p+1}$$

is upper bounded by $O(n^{-1}\log n)$, uniformly for any $k\in\{1,\dots,K\}$. It follows that

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\widehat{\Delta}_k)\overline{X}_i\overline{X}_i^\top+\lambda_n(\tau_{0,k}-\tau_{0,k-1}-\widehat{\tau}_k+\widehat{\tau}_{k-1})\mathbb{E}_{p+1}\right\|_2=O(n^{-1}\log n).$$

In view of (85), we obtain that

$$\max_{k\in\{1,\dots,K\}}\zeta_4(k)=O(n^{-1}\log n),\tag{88}$$

with probability at least $1-O(n^{-2})$. Combining (85)-(88) yields (81).

Proof of Part 2: It follows from Condition (A4) and the definition of the conditional Orlicz norm that

$$\mathrm{E}\left\{\exp\left(\frac{|X^{(j)}|^2}{\omega^2}\right)\right\}=\mathrm{E}\left[\mathrm{E}\left\{\exp\left(\frac{|X^{(j)}|^2}{\omega^2}\right)\middle| A\right\}\right]\le 2,$$

for any $j\in\{1,\dots,p\}$. Without loss of generality, suppose $\omega\ge\log^{-1/2}2$. Then, we have

$$\mathrm{E}\left\{\exp\left(\frac{|\overline{X}^{(j)}|^2}{\omega^2}\right)\right\}=\mathrm{E}\left[\mathrm{E}\left\{\exp\left(\frac{|\overline{X}^{(j)}|^2}{\omega^2}\right)\middle| A\right\}\right]\le 2,$$

for any $j\in\{1,\dots,p+1\}$. As a result, it follows from Bonferroni's inequality and Markov's inequality that

$$\begin{aligned}\Pr\left(\|\overline{X}\|_2>\omega\sqrt{2(p+1)\log n}\right)&\le\sum_{j=1}^{p+1}\Pr(|\overline{X}^{(j)}|>\omega\sqrt{2\log n})\\ \le\ &\sum_{j=1}^{p+1}\mathrm{E}\left\{\exp\left(\frac{|\overline{X}^{(j)}|^2}{\omega^2}\right)\right\}/\exp\left(\frac{2\omega^2\log n}{\omega^2}\right)\le\frac{2(p+1)}{n^2}.\end{aligned}$$

Thus, we obtain

$$\Pr(\mathcal{A}^*) \geq 1 - \frac{2(p+1)}{n^2}, \tag{89}$$

where

$$\mathcal{A}^* = \{\|\overline{X}\|_2 \leq \omega\sqrt{2(p+1)\log n}\}.$$

Consider the event

$$\mathcal{A}_0 = \bigcup_{\mathcal{I}_1,\mathcal{I}_2\in\mathcal{P}_0} \left\{0 < \left|\overline{X}^\top(\theta_{0,\mathcal{I}_1} - \theta_{0,\mathcal{I}_2})\right| \leq \frac{4\sqrt{2(p+1)}c_3\omega\log n}{\sqrt{n\delta_{\min}}}\right\}.$$

By Condition (A5) and Bonferroni's inequality, we have

$$\begin{aligned}\Pr(\mathcal{A}_0) \leq \sum_{\substack{\mathcal{I}_1,\mathcal{I}_2\in\mathcal{P}_0\\ \mathcal{I}_1\neq\mathcal{I}_2}} \Pr\left(0 < \left|\overline{X}^\top(\theta_{0,\mathcal{I}_1} - \theta_{0,\mathcal{I}_2})\right| \leq \frac{4\sqrt{2(p+1)}c_3\omega\log n}{\sqrt{n\delta_{\min}}}\right) \\ \leq K^2\left(\frac{4\sqrt{2(p+1)}c_3\omega\log n}{\sqrt{n\delta_{\min}}}\right)^\gamma.\end{aligned} \tag{90}$$

By the definition of $V^{\pi^*}(\cdot)$, we have

$$V^{\pi^*}(\widehat{d}) = \mathrm{E}\left(\int_{\widehat{d}(X)} \overline{X}^\top\theta_0(a)\pi^*(a; X, \widehat{d}(X))da\right).$$

Notice that the expectation in the above expression is taken with respect to X. Define an interval-valued function $\widehat{d}_0(x) = [\tau_{0,\widehat{\mathbb{K}}(x)-1}, \tau_{0,\widehat{\mathbb{K}}(x)})$ and set $\widehat{\Delta}(x) = \widehat{d}(x) \cap \{\widehat{d}_0(x)\}^c$. It follows that

$$\begin{aligned}V^{\pi^*}(\widehat{d}) &= \mathrm{E}\left(\int_{\widehat{d}_0(X)\cap\widehat{d}(X)} \overline{X}^\top\theta_0(a)\pi^*(a; X, \widehat{d}(X))da\right) + \underbrace{\mathrm{E}\left(\int_{\widehat{\Delta}(X)} \overline{X}^\top\theta_0(a)\pi^*(a; X, \widehat{d}(X))da\right)}_{\chi_1} \\ &= \mathrm{E}\left(\int_{\widehat{d}_0(X)} \overline{X}^\top\theta_0(a)\pi^*(a; X, \widehat{d}(X))da\right) + \chi_1.\end{aligned}$$

Here, the second equality is due to that $\pi^*(a; X, \widehat{d}(X)) = 0$, for any $a \in \{\widehat{d}(X)\}^c$. By (27) and (80), we have

$$|\chi_1| \leq c_0\bar{c}_4\delta_{\min}^{-1}\mathrm{E}\left(\int_{\widehat{\Delta}(X)} \|\overline{X}\|_2 da\right) = c_0\bar{c}_4\delta_{\min}^{-1}\mathrm{E}\|\overline{X}\|_2\lambda(\widehat{\Delta}(X)),$$

where $\lambda(\widehat{\Delta}(X))$ denotes the Lebesgue measure of $\widehat{\Delta}(X)$. Under the event defined in (78), we have $\lambda(\widehat{\Delta}(X)) \leq 2cn^{-1}\log n$, for any realization of X. It follows that

$$|\chi_1| \leq 2cc_0\bar{c}_4\delta_{\min}^{-1}(n^{-1}\log n)\mathrm{E}\|\overline{X}\|_2. \tag{91}$$

By (48), we have

$$\mathrm{E}\|\overline{X}\|_2^2 = \sum_{j=1}^{p+1} \mathrm{E}|\overline{X}^{(j)}|^2 = \sum_{j=1}^{p+1} \mathrm{E}(\mathrm{E}|\overline{X}^{(j)}|^2|A) \le \omega^2(p+1). \tag{92}$$

By Cauchy-Schwarz inequality, this further implies that

$$\mathrm{E}\|\overline{X}\|_2 \le \sqrt{\mathrm{E}\|\overline{X}\|_2^2} \le \omega\sqrt{p+1}.$$

This together with (91) yields

$$|\chi_1| \le 2cc_0\bar{c}_4\omega\sqrt{p+1}\delta_{\min}^{-1}n^{-1}\log n, \tag{93}$$

with probability at least $1-O(n^{-2})$.

Notice that $\theta_0(\cdot)$ is a constant on $\widehat{d}_0(x)$ for any x. It follows that

$$\begin{aligned}\mathrm{E}\left(\int_{\widehat{d}_0(X)} \overline{X}^\top \theta_0(a)\pi^*(a;X,\widehat{d}(X))da\right) &= \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(x)-1},\tau_{0,\widehat{\mathbb{K}}(x)})}\right)\int_{\widehat{d}_0(X)} \pi^*(a;X,\widehat{d}(X))da \\ &= \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(x)-1},\tau_{0,\widehat{\mathbb{K}}(x)})}\right)\int_{\widehat{d}_0(X)\cap\widehat{d}(X)} \pi^*(a;X,\widehat{d}(X))da\end{aligned}$$

$$= \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(x)-1},\tau_{0,\widehat{\mathbb{K}}(x)})}\right)\int_{\widehat{d}(X)} \pi^*(a;X,\widehat{d}(X))da - \chi_2 = \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(x)-1},\tau_{0,\widehat{\mathbb{K}}(x)})}\right) - \chi_2,$$

where

$$\chi_2 = \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(x)-1},\tau_{0,\widehat{\mathbb{K}}(x)})}\right)\int_{\widehat{\Delta}(X)} \pi^*(a;X,\widehat{d}(X))da.$$

Similar to (93), we can show that

$$|\chi_2| = O(n^{-1}\log n),$$

with probability at least $1-O(n^{-2})$. This together with (93) yields (84).

Proof of Part 3: Similar to the definition of $\widehat{\mathbb{K}}$, we define

$$\mathbb{K}_0(x) = \operatorname*{sarg\,max}_{k\in\{1,\dots,K\}} \bar{x}^\top \theta_{0,[\tau_{0,k-1},\tau_{0,k})}. \tag{94}$$

Let

$$\mathbb{K}^*(x) = \left\{k_0 : k_0 = \operatorname*{arg\,max}_{k\in\{1,\dots,K\}} \bar{x}^\top \theta_{0,[\tau_{0,k-1},\tau_{0,k})}\right\},$$

denote the set that consists of all the maximizers. Apparently, $\mathbb{K}_0(x) \in \mathbb{K}^*(x)$, $\forall x \in \mathbb{X}$.

We now claim that

$$\widehat{\mathbb{K}}(X) \in \mathbb{K}^*(X), \tag{95}$$

under the events defined in $\mathcal{A}_0^c \cap \mathcal{A}^*$ and (82). Otherwise, suppose there exists some $k_0 \in \{1, \dots, K\}$ such that

$$\overline{X}^\top \widehat{\theta}_{[\widehat{\tau}_{k_0-1}, \widehat{\tau}_{k_0})} \geq \max_{k \neq k_0} \overline{X}^\top \widehat{\theta}_{[\widehat{\tau}_{k-1}, \widehat{\tau}_k)}, \tag{96}$$

$$\max_{k \neq k_0} \overline{X}^\top \theta_{0,[\tau_{0,k-1}, \tau_{0,k})} > \overline{X}^\top \theta_{0,[\tau_{0,k_0-1}, \tau_{0,k_0})}. \tag{97}$$

Under $\mathcal{A}_0^c$, it follows from (97) that

$$\max_{k \neq k_0} \overline{X}^\top \theta_{0,[\tau_{0,k-1}, \tau_{0,k})} > \overline{X}^\top \theta_{0,[\tau_{0,k_0-1}, \tau_{0,k_0})} + \frac{4\sqrt{2(p+1)} c_3 \omega \log n}{\sqrt{n \delta_{\min}}}. \tag{98}$$

Under the events defined in $\mathcal{A}^*$ and (82), we have

$$\begin{aligned} \max_{k \in \{1,\dots,K\}} |\overline{X}^\top (\widehat{\theta}_{[\widehat{\tau}_{k-1}, \widehat{\tau}_k)} - \theta_{0,[\tau_{0,k-1}, \tau_{0,k})})| &\leq \|\overline{X}\|_2 \max_{k \in \{1,\dots,K\}} \|\widehat{\theta}_{[\widehat{\tau}_{k-1}, \widehat{\tau}_k)} - \theta_{0,[\tau_{0,k-1}, \tau_{0,k})}\|_2 \\ &\leq \frac{2\sqrt{2(p+1)} c_3 \omega \log n}{\sqrt{n \delta_{\min}}}. \end{aligned}$$

This together with (98) yields that

$$\max_{k \neq k_0} \overline{X}^\top \widehat{\theta}_{[\widehat{\tau}_{k-1}, \widehat{\tau}_k)} > \overline{X}^\top \widehat{\theta}_{[\widehat{\tau}_{k_0-1}, \widehat{\tau}_{k_0})}.$$

In view of (96), we have reached an contradiction. Therefore, (95) holds under the events defined in $\mathcal{A}_0^c \cap \mathcal{A}^*$ and (82). When (95) holds, it follows from the definition of $\mathbb{K}^*(\cdot)$ that $\overline{X}^\top \theta_{0,[\widehat{\mathbb{K}}(X)-1, \widehat{\mathbb{K}}(X))} = \overline{X}^\top \theta_{0,[\mathbb{K}_0(X)-1, \mathbb{K}_0(X))}$. Therefore, under the event defined in (82), we have

$$\begin{aligned} \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(X)-1}, \tau_{0,\widehat{\mathbb{K}}(X)})}\right) &= \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(X)-1}, \tau_{0,\widehat{\mathbb{K}}(X)})}\right) \mathbb{I}(\mathcal{A}_0^c \cap \mathcal{A}^*) \qquad (99) \\ + \quad \underbrace{\mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\widehat{\mathbb{K}}(X)-1}, \tau_{0,\widehat{\mathbb{K}}(X)})}\right) \mathbb{I}(\mathcal{A}_0 \cup \mathcal{A}^{*c})}_{\chi_3} &= \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\mathbb{K}_0(X)-1}, \tau_{0,\mathbb{K}_0(X)})}\right) \mathbb{I}(\mathcal{A}_0^c \cap \mathcal{A}^*) \\ + \quad \chi_3 = \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\mathbb{K}_0(X)-1}, \tau_{0,\mathbb{K}_0(X)})}\right) + \chi_3 &- \underbrace{\mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\mathbb{K}_0(X)-1}, \tau_{0,\mathbb{K}_0(X)})}\right) \mathbb{I}(\mathcal{A}_0 \cup \mathcal{A}^{*c})}_{\chi_4}. \end{aligned}$$

Notice that

$$\chi_3 - \chi_4 = \mathrm{E}\overline{X}^\top \left(\theta_{0,[\tau_{0,\widehat{\mathbb{K}}(X)-1}, \tau_{0,\widehat{\mathbb{K}}(X)})} - \theta_{0,[\tau_{0,\mathbb{K}_0(X)-1}, \tau_{0,\mathbb{K}_0(X)})}\right) \mathbb{I}(\mathcal{A}_0 \cup \mathcal{A}^{*c}).$$

Using similar arguments in showing (95), we can show that under the event defined in (82),

$$\overline{X}^\top \left(\theta_{0,[\tau_{0,\widehat{\mathbb{K}}(X)-1}, \tau_{0,\widehat{\mathbb{K}}(X)})} - \theta_{0,[\tau_{0,\mathbb{K}_0(X)-1}, \tau_{0,\mathbb{K}_0(X)})}\right) \neq 0,$$

only when

$$0 < \left|\overline{X}^\top \left(\theta_{0,[\tau_{0,\widehat{\mathbb{K}}(X)-1}, \tau_{0,\widehat{\mathbb{K}}(X)})} - \theta_{0,[\tau_{0,\mathbb{K}_0(X)-1}, \tau_{0,\mathbb{K}_0(X)})}\right)\right| \leq \frac{4\sqrt{2(p+1)} c_3 \omega \log n}{\sqrt{n \delta_{\min}}}.$$

Therefore, under the event defined in (82), we have

$$|\chi_3 - \chi_4| \le \frac{4\sqrt{2(p+1)}c_3\omega \log n}{\sqrt{n\delta_{\min}}}\Pr(\mathcal{A}_0 \cup \mathcal{A}^{*c}).$$

It follows from (89) and (90), we have

$$|\chi_3 - \chi_4| \le \frac{4\sqrt{2(p+1)}c_3\omega \log n}{\sqrt{n\delta_{\min}}}\left\{\frac{2(p+1)}{n^2} + K^2\left(\frac{4\sqrt{2(p+1)}c_3\omega \log n}{\sqrt{n\delta_{\min}}}\right)^{\gamma}\right\}.$$

For sufficiently large n, this together with (84) and (99) implies that we have with probability at least $1 - O(n^{-2})$,

$$V^{\pi^*}(\widehat{d}) \ge \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\mathbb{K}_0(X)-1},\tau_{0,\mathbb{K}_0(X)})}\right) - O(1)(n^{-1}\log n + n^{-(1+\gamma)/2}\log^{1+\gamma} n),$$

for some positive constant $O(1)$. The proof is hence completed by noting that

$$V^{opt} = \mathrm{E}\left(\overline{X}^\top \theta_{0,[\tau_{0,\mathbb{K}_0(X)-1},\tau_{0,\mathbb{K}_0(X)})}\right).$$

B.7 Proof of Theorem 3

We first introduce some technical lemmas. We remark that the key ingredient of the proof lies in Lemma 5, which establishes a uniform upper bound on the mean squared error of $\widehat{q}_{\mathcal{I}}$. Proofs of these lemmas can be found in Sections E.1 - E.3 of Cai et al. (2021)[2] and we omit them for brevity. The rest of the proof can be similarly proven as Theorem 1. Specifically, we first show the consistency of the estimated change point locations. We then derive the rate of convergence of the estimated change point locations and the estimated outcome regression function.

Lemma 5 *Assume conditions in Theorem 3 are satisfied. Then there exists some constant* $\bar{C} > 0$ *such that the following holds with probability at least* $1 - O(n^{-2})$*: For any* $\mathcal{I} \in \mathfrak{I}(m)$ *and* $|\mathcal{I}| \ge c\gamma_n$*,*

$$E|q_{\mathcal{I},0}(X) - \widehat{q}_{\mathcal{I}}(X)|^2 \le \bar{C}(n|\mathcal{I}|)^{-2\beta/(2\beta+p)}\log^8 n, \tag{100}$$

where $q_{\mathcal{I},0} = E(Y|A \in \mathcal{I}, X)$ *for any interval* $\mathcal{I}$*.* ■

Lemma 6 *Assume conditions in Theorem 3 are satisfied. Then there exists some constant* $\bar{C} > 0$ *such that the followings hold with probability at least* $1 - O(n^{-2})$*: For any* $\mathcal{I} \in \mathfrak{I}(m)$ *and* $|\mathcal{I}| \ge c\gamma_n$*,*

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\left|\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})\{Y_i - q_{\mathcal{I},0}(X_i)\}\{\widehat{q}_{\mathcal{I}}(X_i) - q_{\mathcal{I},0}(X_i)\}\right| \le \bar{C}(n|\mathcal{I}|)^{p/(2\beta+p)}\log^8 n,$$

for any $\mathcal{I} \in \mathfrak{I}(m)$ *such that* $|\mathcal{I}| \ge c\gamma_n$ *for any positive constant* $c > 0$*.* ■

Lemma 7 *Under the conditions in Theorem 3, the following events occur with probability at least* $1 - O(n^{-2})$*: there exists some constant* $C > 0$ *such that* $\min_{\mathcal{I}\in\widehat{\mathcal{P}}}|\mathcal{I}| \ge C\gamma_n$*.* ■

[2] See `https://openreview.net/attachment?id=rvKD3iqtBdk&name=supplementary_material`

We next show the consistency of the estimated change-point locations. Using similar arguments in proving (31), we can show that

$$|\widehat{\mathcal{P}}| \le C_0 \gamma_n^{-1}, \tag{101}$$

for sufficiently large n and some constant $C_0 > 0$.

Notice that

$$\begin{aligned}
\sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \widehat{q}_{\mathcal{I}}(X_i)\}^2 \ge & \underbrace{\sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Y_i - q_{\mathcal{I},0}(X_i)\}^2}_{\eta_1^*} \\
& + \sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{\widehat{q}_{\mathcal{I}}(X_i) - q_{\mathcal{I},0}(X_i)\}^2 \\
& -2 \sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \left| \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Y_i - q_{\mathcal{I},0}(X_i)\}\{\widehat{q}_{\mathcal{I}}(X_i) - q_{\mathcal{I},0}(X_i)\} \right|.
\end{aligned}$$

The second line is non-negative. Under Lemmas 6 and 7, the third line is lower bounded by $-C_1 \sum_{\mathcal{I}\in\widehat{\mathcal{P}}} (n|\mathcal{I}|)^{p/(2\beta+p)} \log^8 n$ for some constant $C_1 > 0$. By Hölder's inequality, it can be further lower bounded by $-C_1 |\widehat{\mathcal{P}}|^{2\beta/(2\beta+p)} n^{p/(2\beta+p)} \log^8 n$. By (101) and the given condition on γ_n, the third line is $o(n)$. It follows that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \widehat{q}_{\mathcal{I}}(X_i)\}^2 \ge \eta_1^* + o(n), \tag{102}$$

with probability at least $1 - O(n^{-2})$.

Similar to (24) and (25), we can show that the following events occur with probability at least $1 - O(n^{-2})$,

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Y_i - Q(X_i, A_i)\}\{Q(X_i, A_i) - q_{\mathcal{I},0}(X_i)\} \right| \\
& \quad \le c_0 \left[n^{-1/2} \sqrt{\mathrm{E}\mathbb{I}(A \in \mathcal{I})\{Q(X, A) - q_{\mathcal{I},0}(X)\}^2 \log n} + n^{-1} \log n \right], \\
& \left| \frac{1}{n} \sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{Q(X_i, A_i) - q_{\mathcal{I},0}(X_i)\}^2 - \mathrm{E}\mathbb{I}(A \in \mathcal{I})|Q(X, A) - q_{\mathcal{I},0}(X)|^2 \right| \\
& \quad \le c_0 \left[n^{-1/2} \sqrt{\mathrm{E}\mathbb{I}(A \in \mathcal{I})\{Q(X, A) - q_{\mathcal{I},0}(X)\}^2 \log n} + n^{-1} \log n \right],
\end{aligned}$$

for some constant $c_0 > 0$ and any $\mathcal{I}$. The two upper bounds are $o(1)$. Similar to (35), we can show that

$$\eta_1^* = \sum_{i=1}^n |Y_i - Q(X_i, A_i)|^2 + n \sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \mathrm{E}\mathbb{I}(A \in \mathcal{I})|Q(X, A) - q_{\mathcal{I},0}(X)|^2 + o(n),$$

with probability at least $1-O(n^{-2})$. It follows from (102) that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2 \ge \underbrace{\sum_{i=1}^n |Y_i-Q(X_i,A_i)|^2}_{\eta_2^*} + n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathrm{E}\mathbb{I}(A\in\mathcal{I})|Q(X,A)-q_{\mathcal{I},0}(X)|^2+o(n), \tag{103}$$

with probability at least $1-O(n^{-2})$.

Let us consider η_2^*. We observe that

$$\eta_2^*=\sum_{\mathcal{I}\in\mathcal{P}_0}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})|Y_i-q_{\mathcal{I},0}(X_i)|^2.$$

By the uniform approximation property of DNN, there exists some $q_{\mathcal{I}}^*\in\mathcal{Q}_{\mathcal{I}}$ such that

$$\sum_{i=1}^n |q_{\mathcal{I},0}(X_i)-q_{\mathcal{I}}^*(X_i)|^2\propto n(n|\mathcal{I}|)^{-2\beta/(2\beta+p)}.$$

See Part 1 of the proof of Lemma 5 for details. Similar to (24) and (25), we can show that the following events occur with probability at least $1-O(n^{-2})$,

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\{Y_i-q_{\mathcal{I},0}(X_i)\}\{q_{\mathcal{I},0}(X_i)-q_{\mathcal{I}}^*(X_i)\}\right|\le \frac{c_0\sqrt{|\mathcal{I}|\log n}}{\sqrt{n}}(n|\mathcal{I}|)^{-\beta/(2\beta+p)},$$

for some constant $c_0>0$ and any $\mathcal{I}\in\mathcal{P}_0$. It follows that

$$\eta_2^*-\sum_{\mathcal{I}\in\mathcal{P}_0}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})|Y_i-q_{\mathcal{I}}^*(X_i)|^2\ge -\sum_{\mathcal{I}\in\mathcal{P}_0}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})|q_{\mathcal{I},0}(X_i)-q_{\mathcal{I}}^*(X_i)|^2 -2\sum_{\mathcal{I}\in\mathcal{P}_0}\left|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\{Y_i-q_{\mathcal{I},0}(X_i)\}\{q_{\mathcal{I},0}(X_i)-q_{\mathcal{I}}^*(X_i)\}\right|\ge -\bar{c}n^{p/(2\beta+p)},$$

for some constant $\bar{c}>0$. This together with (103) yields that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2\ge\sum_{\mathcal{I}\in\mathcal{P}_0}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})|Y_i-q_{\mathcal{I}}^*(X_i)|^2 + n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathrm{E}\mathbb{I}(A\in\mathcal{I})|Q(X,A)-q_{\mathcal{I},0}(X)|^2+o(n)+O(n^{p/(2\beta+p)}),$$

with probability at least $1-O(n^{-2})$.

Next, using similar arguments in proving (39), we can show that there exist a partition $\mathcal{P}^*\in\mathcal{B}(m)$ and a set of functions $\{q_{\mathcal{I}}^{**}:\mathcal{I}\in\mathcal{P}^*\}$ with $|\mathcal{P}^*|=|\mathcal{P}_0|$ such that

$$\sum_{\mathcal{I}\in\mathcal{P}_0}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})|Y_i-q_{\mathcal{I}}^{**}(X_i)|^2\ge\sum_{\mathcal{I}\in\mathcal{P}^*}\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})|Y_i-q_{\mathcal{I}}^{**}(X_i)|^2+O(1).$$

It follows that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2 \geq \sum_{\mathcal{I}\in\mathcal{P}^*}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})|Y_i-q_{\mathcal{I}}^{**}(X_i)|^2 \tag{104}$$
$$+n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathrm{E}\mathbb{I}(A\in\mathcal{I})|Q(X,A)-q_{\mathcal{I},0}(X)|^2+o(n)+O(n^{p/(2\beta+p)}),$$

with probability at least $1-O(n^{-2})$. Since

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2+n\gamma_n|\widehat{\mathcal{P}}| \tag{105}$$
$$\leq \sum_{\mathcal{I}\in\mathcal{P}^*}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})|Y_i-q_{\mathcal{I}}^{**}(X_i)|^2+n\gamma_n|\mathcal{P}_0|,$$

and that $\gamma_n\to 0$, we obtain that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathrm{E}\mathbb{I}(A\in\mathcal{I})|Q(X,A)-q_{\mathcal{I},0}(X)|^2=o(1).$$

Under the condition that $q_{\mathcal{I}_1,0}\neq q_{\mathcal{I}_2,0}$ for any adjacent $\mathcal{I}_1,\mathcal{I}_2\in\mathcal{P}_0$, we have $\mathrm{E}|q_{\mathcal{I}_1,0}(X)-q_{\mathcal{I}_2,0}(X)|^2>0$. Using similar arguments in the Part 1 of the proof of Theorem 1, we obtain that $\max_{\tau\in J(\mathcal{P}_0)}\min_{\hat{\tau}\in J(\widehat{\mathcal{P}})}|\hat{\tau}-\tau|\leq\delta$ for any constant $\delta>0$. This further implies that $|\widehat{\mathcal{P}}|\geq|\mathcal{P}_0|$.

We next derive the rate of convergence of the estimated change point locations and the estimated outcome regression function. Similar to (104), with a more refined analysis (see e.g., Step 2 of the proof of Theorem 1), we obtain that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2 \geq \sum_{\mathcal{I}\in\mathcal{P}^*}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})|Y_i-q_{\mathcal{I}}^{**}(X_i)|^2$$
$$+n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathrm{E}\mathbb{I}(A\in\mathcal{I})|Q(X,A)-q_{\mathcal{I},0}(X)|^2-C_1|\widehat{\mathcal{P}}|^{2\beta/(2\beta+p)}n^{p/(2\beta+p)}\log^8 n+O(n^{p/(2\beta+p)}),$$

with probability at least $1-O(n^{-2})$. This together with (105) yields that

$$n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathrm{E}\mathbb{I}(A\in\mathcal{I})|Q(X,A)-q_{\mathcal{I},0}(X)|^2\leq C_1|\widehat{\mathcal{P}}|^{2\beta/(2\beta+p)}n^{p/(2\beta+p)}\log^8 n$$
$$+O(n^{p/(2\beta+p)})+n\gamma_n(|\mathcal{P}_0|-|\widehat{\mathcal{P}}|).$$

Under the given condition on γ_n, we obtain that $|\widehat{\mathcal{P}}|\leq|\mathcal{P}_0|$. Combining this together with $|\widehat{\mathcal{P}}|\geq|\mathcal{P}_0|$, we obtain that $|\widehat{\mathcal{P}}|=|\mathcal{P}_0|$. This proves the results in (i).

Consequently, we obtain that

$$n\sum_{\mathcal{I}\in\widehat{\mathcal{P}}}\mathrm{E}\mathbb{I}(A\in\mathcal{I})|Q(X,A)-q_{\mathcal{I},0}(X)|^2=O(n^{p/(2\beta+p)}\log^8 n),$$

As such, we have that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \mathrm{E}\mathbb{I}(A\in\mathcal{I})|Q(X,A)-q_{\mathcal{I},0}(X)|^2 = O(n^{-2\beta/(2\beta+p)}\log^8 n),$$

This together with Lemma 5 proves the result in (iii). Using similar arguments in Part 2 of the proof of Theorem 1, we can show the result in (ii) holds. This completes the proof.

B.8 Proof of Theorem 4

The proof of Theorem 4 is similar to that of Theorem 2. We provide the outline as below and omit the duplicated arguments for brevity.

Under the events defined in Theorem 3, we have $\widehat{K}=K$, and

$$\max_{k\in\{1,\dots,K-1\}} |\widehat{\tau}_k-\tau_{0,k}| \le cn^{-2\beta/(2\beta+p)}\log^8 n, \tag{106}$$

for some constant $c>0$. By similar arguments in the proof of Theorem 2, there exists some constant $\bar{C}_4>0$ such that

$$\pi^*(a;x,\widehat{d}(x)) \le \bar{C}_4\delta_{\min}^{-1}, \quad \forall a\in[0,1], x\in\mathbb{X}. \tag{107}$$

The rest of our proof is divided into two parts. In the first part, we focus on proving

$$V^{\pi^*}(\widehat{d}) \ge \mathrm{E}\left(q_{[\tau_{0,\widehat{\mathbb{K}}(x)-1},\tau_{0,\widehat{\mathbb{K}}(x)})}(X)\right) - O(1)n^{-2\beta/(2\beta+p)}\log^8 n, \tag{108}$$

with probability at least $1-O(n^{-2})$, where $O(1)$ denotes some positive constant.

In Part 2, we provide an upper bound for

$$V^{opt}-\mathrm{E}\left(q_{[\tau_{0,\widehat{\mathbb{K}}(x)-1},\tau_{0,\widehat{\mathbb{K}}(x)})}(X)\right).$$

This together with (108) yields the desired results.

Proof of Part 1: Recall the integer-valued function

$$\widehat{\mathbb{K}}(x) = \operatorname*{sarg\,max}_{k\in\{1,\dots,K\}} \widehat{q}_{[\widehat{\tau}_{k-1},\widehat{\tau}_k)}(x), \tag{109}$$

where sarg max denotes the smallest maximizer when the argmax is not unique. Similarly, we have $\widehat{\mathbb{K}}(x)=k$ if $\widehat{d}(x)=[\widehat{\tau}_{k-1},\widehat{\tau}_k)$ for some integer k such that $1\le k\le K-1$, and set $\widehat{\mathbb{K}}(x)=K$ if $\widehat{d}(x)=[\widehat{\tau}_{K-1},1]$.

Let $\widehat{\Delta}_k=[\widehat{\tau}_{k-1},\widehat{\tau}_k)\cup[\tau_{0,k-1},\tau_{0,k})^c+[\widehat{\tau}_{k-1},\widehat{\tau}_k)^c\cup[\tau_{0,k-1},\tau_{0,k})$. Using similar arguments in the proof of Theorem 2, we have

$$V^{\pi^*}(\widehat{d}) = \mathrm{E}\left(\int_{\widehat{d}_0(X)\cap\widehat{d}(X)} Q(X,a)\pi^*(a;X,\widehat{d}(X))da\right) + \underbrace{\mathrm{E}\left(\int_{\widehat{\Delta}(X)} Q(X,a)\pi^*(a;X,\widehat{d}(X))da\right)}_{\chi_1^*}$$

$$= \mathrm{E}\left(\int_{\widehat{d}_0(X)} Q(X,a)\pi^*(a;X,\widehat{d}(X))da\right) + \chi_1^*,$$

where $\widehat{d}_0(x) = [\tau_{0,\widehat{\mathbb{K}}(x)-1}, \tau_{0,\widehat{\mathbb{K}}(x)})$ and $\widehat{\Delta}(x) = \widehat{d}(x) \cap \{\widehat{d}_0(x)\}^c$.

By (107) and the assumption that Y is bounded, we have

$$|\chi_1^*| \le c_0 \bar{C}_4 \delta_{\min}^{-1} \lambda(\widehat{\Delta}(X)),$$

where $\lambda(\widehat{\Delta}(X))$ denotes the Lebesgue measure of $\widehat{\Delta}(X)$. Under the event defined in (106), we have $\lambda(\widehat{\Delta}(X)) \le 2cn^{-2\beta/(2\beta+p)} \log^8 n$, for any realization of X. It follows that

$$|\chi_1^*| \le \bar{C}_0 \delta_{\min}^{-1} n^{-2\beta/(2\beta+p)} \log^8 n, \tag{110}$$

for some constant $\bar{C}_0$ with probability at least $1 - O(n^{-2})$.

Using similar arguments in the proof of Theorem 2, we have

$$\mathrm{E}\left(\int_{\widehat{d}_0(X)} Q(X,a)\pi^*(a; X, \widehat{d}(X)) da \right) = \mathrm{E}\left(q_{[\tau_{0,\widehat{\mathbb{K}}(x)-1}, \tau_{0,\widehat{\mathbb{K}}(x)})}(X) \right) - \chi_2^*,$$

where

$$\chi_2^* = \mathrm{E}\left(q_{[\tau_{0,\widehat{\mathbb{K}}(x)-1}, \tau_{0,\widehat{\mathbb{K}}(x)}),0}(X) \right) \int_{\widehat{\Delta}(X)} \pi^*(a; X, \widehat{d}(X)) da.$$

Similar to (110), we can show that

$$|\chi_2^*| = O(n^{-2\beta/(2\beta+p)} \log^8 n),$$

with probability at least $1 - O(n^{-2})$. This together with (110) yields (108).

Proof of Part 2: Let $\epsilon_n = \bar{C}_1 (n\delta_{\min})^{-2\beta/\{(2\beta+p)(2+\gamma)\}} \log^{8/(2+\gamma)} n$ for some constant $\bar{C}_1$. Define an event

$$\mathcal{A}_\epsilon = \bigcup_k \left\{ |q_{[\tau_{0,k-1},\tau_{0,k})}(X) - \widehat{q}_{[\widehat{\tau}_{k-1},\widehat{\tau}_k)}(X)| \le \epsilon_n \right\}.$$

Based on Lemma 5, by Markov's inequality, we can show that there exists some constant $\bar{c} > 0$ such that

$$\begin{aligned} & \Pr\{|q_{[\tau_{0,k-1},\tau_{0,k}),0}(X) - \widehat{q}_{[\widehat{\tau}_{k-1},\widehat{\tau}_k)}(X)| > \epsilon_n\} \\ \le \quad & \bar{C}_2 (n\delta_{\min})^{-2\beta(1+\gamma)/\{(2\beta+p)(2+\gamma)\}} \log^{8(1+\gamma)/(2+\gamma)} n, \forall k \in \{1, \dots, K\}, \end{aligned} \tag{111}$$

with probability at least $1 - O(n^{-2})$ for some constant $\bar{C}_2$. Thus, by Bonferroni's inequality, we have

$$\Pr\{\mathcal{A}_\epsilon^c\} \le \bar{C}_3 (n\delta_{\min})^{-2\beta(1+\gamma)/\{(2\beta+p)(2+\gamma)\}} \log^{8(1+\gamma)/(2+\gamma)} n \tag{112}$$

holds with probability at least $1 - O(n^{-2})$ for some constant $\bar{C}_3$.

Consider the event

$$\mathcal{A}_0 = \bigcup_{\mathcal{I}_1, \mathcal{I}_2 \in \mathcal{P}_0} \{0 < |q_{\mathcal{I}_1,0}(X) - q_{\mathcal{I}_2,0}(X)| \le 2\epsilon_n\}.$$

By Condition (A5) and Bonferroni's inequality, we have

$$\Pr(\mathcal{A}_0) \le \sum_{\substack{\mathcal{I}_1,\mathcal{I}_2\in\mathcal{P}_0\\ \mathcal{I}_1\neq\mathcal{I}_2}} \Pr\left(0 < |q_{\mathcal{I}_1,0}(X) - q_{\mathcal{I}_2,0}(X)| \le 2\epsilon_n\right) \le K^2\left(2\epsilon_n\right)^{\gamma}. \tag{113}$$

Similar to the definition of $\widehat{\mathbb{K}}$, we define

$$\mathbb{K}_0(x) = \operatorname*{sarg\,max}_{k\in\{1,\dots,K\}} q_{[\tau_{0,k-1},\tau_{0,k}),0}(x). \tag{114}$$

Let

$$\mathbb{K}^*(x) = \left\{ k_0 : k_0 = \operatorname*{arg\,max}_{k\in\{1,\dots,K\}} q_{[\tau_{0,k-1},\tau_{0,k}),0}(x) \right\},$$

denote the set that consists of all the maximizers. Apparently, $\mathbb{K}_0(x) \in \mathbb{K}^*(x)$, $\forall x \in \mathbb{X}$.

We now claim that

$$\widehat{\mathbb{K}}(X) \in \mathbb{K}^*(X), \tag{115}$$

under the events defined in $\mathcal{A}_0^c$ and $\mathcal{A}_\epsilon$. Otherwise, suppose there exists some $k_0 \in \{1,\dots,K\}$ such that

$$\widehat{q}_{[\widehat{\tau}_{k_0-1},\widehat{\tau}_{k_0})}(X) \ge \max_{k\neq k_0} \widehat{q}_{[\widehat{\tau}_{k-1},\widehat{\tau}_k)}(X), \tag{116}$$

$$\max_{k\neq k_0} q_{[\tau_{0,k-1},\tau_{0,k}),0}(X) > q_{[\tau_{0,k_0-1},\tau_{0,k_0}),0}(X). \tag{117}$$

Under $\mathcal{A}_0^c$, it follows from (117) that

$$\max_{k\neq k_0} q_{[\tau_{0,k-1},\tau_{0,k}),0}(X) > q_{[\tau_{0,k_0-1},\tau_{0,k_0}),0} + 2\epsilon_n. \tag{118}$$

Under the event $\mathcal{A}_\epsilon$, we have

$$\max_{k\in\{1,\dots,K\}} |\widehat{q}_{[\widehat{\tau}_{k-1},\widehat{\tau}_k)}(X) - q_{[\tau_{0,k-1},\tau_{0,k}),0}(X)| \le \epsilon_n.$$

This together with (118) yields that

$$\max_{k\neq k_0} \widehat{q}_{[\widehat{\tau}_{k-1},\widehat{\tau}_k)}(X) > \widehat{q}_{[\widehat{\tau}_{k_0-1},\widehat{\tau}_{k_0})}(X).$$

In view of (116), we have reached a contradiction. Therefore, (115) holds under the events defined in $\mathcal{A}_0^c$ and $\mathcal{A}_\epsilon$.

By the definition of $\mathbb{K}^*(\cdot)$ that $q_{[\tau_{0,\widehat{\mathbb{K}}(X)-1},\tau_{0,\widehat{\mathbb{K}}(X)}),0}(X) = q_{[\tau_{0,\mathbb{K}_0(X)-1},\tau_{0,\mathbb{K}_0(X)}),0}(X)$ when (115) holds. Using the similar augments in (99), we have

$$\mathrm{E}\left(q_{[\tau_{0,\widehat{\mathbb{K}}(x)-1},\tau_{0,\widehat{\mathbb{K}}(x)}),0}(X)\right) = \mathrm{E}\left(q_{[\tau_{0,\mathbb{K}_0(X)-1},\tau_{0,\mathbb{K}_0(X)}),0}(X)\right) + \chi_3 + \chi_4, \tag{119}$$

where

$$\chi_3 = \mathrm{E}\left(q_{[\tau_{0,\widehat{\mathbb{K}}(X)-1},\tau_{0,\widehat{\mathbb{K}}(X)}),0}(X) - q_{[\tau_{0,\mathbb{K}_0(X)-1},\tau_{0,\mathbb{K}_0(X)}),0}(X)\right)\mathbb{I}(\mathcal{A}_0)\mathbb{I}(\mathcal{A}_\epsilon),$$

and

$$\chi_4 = \mathrm{E}\left(q_{[\tau_{0,\widehat{\mathbb{K}}(X)-1},\tau_{0,\widehat{\mathbb{K}}(X)}),0}(X) - q_{[\tau_{0,\mathbb{K}_0(X)-1},\tau_{0,\mathbb{K}_0(X)}),0}(X)\right)\mathbb{I}(\mathcal{A}_\epsilon^c),$$

Therefore, under the event $\mathcal{A}_\epsilon$, it follows from (113) that

$$|\chi_3| \leq K^2 \left(2\epsilon_n\right)^{\gamma+1}. \tag{120}$$

Similarly, by Condition (A7) and the outcome is bounded, following Markov's inequality, we have

$$|\chi_4| \quad \leq \bar{C}_3 \mathrm{Pr}\{\mathcal{A}_\epsilon^c\} \tag{121}$$

Based on (112) and $\epsilon_n = \bar{C}_1(n\delta_{\min})^{-2\beta/\{(2\beta+p)(2+\gamma)\}}\log^{8/(2+\gamma)} n$, for sufficiently large n, the above (121) and (120) together with (108) and (119) implies that we have with probability at least $1 - O(n^{-2})$,

$$V^{\pi^*}(\widehat{d}) \geq V^{opt} - O(1)(n^{-\frac{2\beta}{2\beta+p}}\log^8 n + n^{-\frac{2\beta(1+\gamma)}{(2\beta+p)(2+\gamma)}}\log^{\frac{8+8\gamma}{2+\gamma}} n),$$

for some positive constant $O(1)$. The proof is hence completed.

B.9 Proof of Theorem 5

We focus on proving Theorem 5 (ii) when conditions in Theorem 4 are satisfied with $4\beta(1+\gamma) > (2\beta+p)(2+\gamma)$, where D-JIL is applied. Since the piecewise linear case requires weaker conditions (when conditions in Theorem 2 are satisfied), one can similarly derive the asymptotic normality of $\widehat{V}$ under L-JIL.

We present an outline of the proof first, which can be divided into two parts. Define $d_0(x) = \arg\max_{\mathcal{I}\in\mathcal{P}_0} q_{\mathcal{I},0}(x)$, $\forall x \in \mathbb{X}$. Under the given conditions, the maximizers $d_0(X_i)$'s are almost surely unique. By the definition of $\widehat{\mathbb{K}}(\cdot)$ in (109), we have

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^n \left[\frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)}\{Y_i - \widehat{q}_{[\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)-1},\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)})}(X_i)\} + \widehat{q}_{[\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)-1},\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)})}(X_i)\right].$$

Given $\mathbb{K}_0(\cdot)$ defined in (114), the above value estimator can be decomposed by

$$\widehat{V} = \widehat{V}_1 + \underbrace{\frac{1}{n}\sum_{i=1}^n \left[\left\{\frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)} - 1\right\}\left\{q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i) - \widehat{q}_{[\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)-1},\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)})}(X_i)\right\}\right]}_{\eta_7},$$

where

$$\widehat{V}_1 = \frac{1}{n}\sum_{i=1}^n \left[\frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)}\{Y_i - q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i)\} + q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i)\right].$$

In Part 1, we first establish the following result that

$$\eta_7 = o_p(n^{-1/2}). \tag{122}$$

This implies

$$\widehat{V} = \widehat{V}_1 + o_p(n^{-1/2}). \tag{123}$$

In the second step, we further decompose $\widehat{V}_1$ as

$$\widehat{V}_1 \quad = \widehat{V}_2 + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)} - \frac{\mathbb{I}\{A_i \in d_0(X_i)\}}{e(d_0(X_i)|X_i)}\right\}\{Y_i - q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i)\}\right]}_{\eta_8},$$

where

$$\widehat{V}_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}\{A_i \in d_0(X_i)\}}{e(d_0(X_i)|X_i)}\{Y_i - q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i)\} + q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i)\right].$$

We focus on proving

$$\eta_8 = o_p(n^{-1/2}). \tag{124}$$

This together with (123) leads to

$$\widehat{V} = \widehat{V}_2 + o_p(n^{-1/2}). \tag{125}$$

Combing the results in the first two steps, it follows from the definition of $d_0(\cdot)$ that

$$\widehat{V} = \quad \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{I}\in\mathcal{P}_0}\mathbb{I}(\mathcal{I} = d_0(X_i))\left[\frac{\mathbb{I}\{A_i \in d_0(X_i)\}}{e(d_0(X_i)|X_i)}\{Y_i - q_{\mathcal{I},0}(X_i)\} + q_{\mathcal{I},0}(X_i)\right] + o_p(n^{-1/2}),$$

almost surely. Notice that the first term at RHS corresponds to a sum of i.i.d random variables. Hence, based on Lindeberg-Feller central limit theorem, one can show the asymptotic normality result of the value estimator under the proposed I2DR.

Proof of Part 1: We aim to show (122). Toward that end, we define

$$\widehat{V}_3 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)}\{Y_i - q_{[\tau_{0,\widehat{\mathbb{K}}(X_i)-1},\tau_{0,\widehat{\mathbb{K}}(X_i)}),0}(X_i)\} + q_{[\tau_{0,\widehat{\mathbb{K}}(X_i)-1},\tau_{0,\widehat{\mathbb{K}}(X_i)}),0}(X_i)\right].$$

The difference $|\eta_7|$ can be upper bounded by $|\widehat{V}_1 - \widehat{V}_3| + |\widehat{V} - \widehat{V}_3|$. Consider $|\widehat{V}_1 - \widehat{V}_3|$ first. Under the given conditions, the term $\left\{\frac{\mathbb{I}\{A_i\in\widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)} - 1\right\}$ is bounded, it suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n}\left|q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i) - q_{[\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)-1},\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)}),0}(X_i)\right| = o_p(n^{-1/2}),$$

where $\mathbb{K}_0(\cdot)$ and $\widehat{\mathbb{K}}(\cdot)$ are defined in (114) and (109), respectively. Under the margin-type condition, the above expression can be proven using similar arguments in the proof of Theorem 4. We omit the details to save space.

It remains to show $|\widehat{V} - \widehat{V}_3| = o_p(n^{1/2})$. Notice that $|\widehat{V} - \widehat{V}_3|$ can be further upper bounded by

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{e(\widehat{d}(X_i)|X_i)} - 1 \right\} \left\{ q_{[\tau_{0,\widehat{\mathbb{K}}(X_i)-1},\tau_{0,\widehat{\mathbb{K}}(X_i)}),0}(X_i) - \widehat{q}_{[\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)-1},\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)})}(X_i) \right\} \right|$$
$$+ \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{e(\widehat{d}(X_i)|X_i)} - \frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)} \right\} \left\{ q_{[\tau_{0,\widehat{\mathbb{K}}(X_i)-1},\tau_{0,\widehat{\mathbb{K}}(X_i)}),0}(X_i) - \widehat{q}_{[\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)-1},\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)})}(X_i) \right\} \right|.$$

Consider the first line. Notice that it can be represented by

$$\left| \frac{1}{n} \sum_{\mathcal{I} \in \widehat{\mathcal{P}}} \sum_{i=1}^{n} \left\{ \frac{\mathbb{I}\{A_i \in \mathcal{I}\}}{e(\mathcal{I}|X_i)} - 1 \right\} \left\{ q_{\mathcal{I},0}(X_i) - \widehat{q}_{\mathcal{I}}(X_i) \right\} \mathbb{I}(\mathcal{I} = \widehat{d}(X_i)) \right|.$$

Since the number of intervals in $\widehat{\mathcal{P}}$ is finite with probability tending to 1 (see Results (i) in Theorem 1), to show the above expression is $o_p(n^{-1/2})$, it suffices to show

$$\sup_{\mathcal{I} \in \mathfrak{I}(m)} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\mathbb{I}\{A_i \in \mathcal{I}\}}{e(\mathcal{I}|X_i)} - 1 \right\} \left\{ q_{\mathcal{I},0}(X_i) - \widehat{q}_{\mathcal{I}}(X_i) \right\} \mathbb{I}(\mathcal{I} = \widehat{d}(X_i)) \right| = o_p(n^{-1/2}).$$

The key observation is that, by Corollary A.1 of Chernozhukov et al. (2014), the above empirical sum forms a VC-type class. Using similar arguments in bounding the stochastic error in Step 2 of the proof of Lemma 5, we can show the above assertion holds.

To bound the second line, notice that by Cauchy-Schwarz inequality, it is smaller than or equal to the square root of

$$\underbrace{\frac{1}{n} \sum_{i=1}^{n} \left| \frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{e(\widehat{d}(X_i)|X_i)} - \frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)} \right|^2}_{\eta_7^{(1)}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} |q_{[\tau_{0,\widehat{\mathbb{K}}(X_i)-1},\tau_{0,\widehat{\mathbb{K}}(X_i)}),0}(X_i) - \widehat{q}_{[\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)-1},\widehat{\tau}_{\widehat{\mathbb{K}}(X_i)})}(X_i)|^2}_{\eta_7^{(2)}}$$

Using similar arguments in establishing the uniform convergence rate of $\widehat{q}_{\mathcal{I}}$, we can show that $\eta_7^{(2)} = o_p(n^{-c})$ for some $c > 1/2$. To prove the second line is $o_p(n^{-1/2})$, it remains to show $\eta_7^{(1)} = O_p(n^{-1/2} \log n)$. Under the positivity assumption on e and $\widehat{e}$, it suffices to show

$$\frac{1}{n} \sum_{i=1}^{n} |e(\widehat{d}(X_i)|X_i) - \widehat{e}(\widehat{d}(X_i)|X_i)|^2 = O_p(n^{1/2} \log n). \tag{126}$$

The left-hand-side can be further upper bounded by

$$\frac{1}{n} \sum_{\mathcal{I} \in \widehat{\mathcal{P}}} \sum_{i=1}^{n} |e(\mathcal{I}|X_i) - \widehat{e}(\mathcal{I}|X_i)|^2$$
$$\leq \sum_{\mathcal{I} \in \widehat{\mathcal{P}}} \mathrm{E}|e(\mathcal{I}|X) - \widehat{e}(\mathcal{I}|X)|^2 + \sum_{\mathcal{I} \in \widehat{\mathcal{P}}} \left[\frac{1}{n} \sum_{i=1}^{n} |e(\mathcal{I}|X_i) - \widehat{e}(\mathcal{I}|X_i)|^2 - \mathrm{E}|e(\mathcal{I}|X) - \widehat{e}(\mathcal{I}|X)|^2 \right].$$

The first term on the second line is $O_p(n^{-1/2})$ under Condition (A8) and the fact that $|\widehat{\mathcal{P}}| = O(1)$ with probability tending to 1. To prove (126), by the boundedness of $|\widehat{\mathcal{P}}|$, it suffices to show the supremum of the empirical process term

$$\sup_{\mathcal{I}\in\mathfrak{I}(m)} \left[\frac{1}{n}\sum_{i=1}^{n} |e(\mathcal{I}|X_i) - \widehat{e}(\mathcal{I}|X_i)|^2 - \mathrm{E}|e(\mathcal{I}|X) - \widehat{e}(\mathcal{I}|X)|^2\right] = O_p(n^{-1/2}\log n).$$

Under Condition (A8), this can be proven in a similar manner as Step 2 of the proof of Lemma 5. We omit the details to save space.

Proof of Part 2: We next focus on proving (124). We notice that $|\eta_8|$ can be upper bounded by

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{\widehat{e}(\widehat{d}(X_i)|X_i)} - \frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{e(\widehat{d}(X_i)|X_i)}\right\}\left\{Y_i - q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i)\right\}\right]\right|$$
$$+\left|\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\frac{\mathbb{I}\{A_i \in d_0(X_i)\}}{e(d_0(X_i)|X_i)} - \frac{\mathbb{I}\{A_i \in \widehat{d}(X_i)\}}{e(\widehat{d}(X_i)|X_i)}\right\}\left\{Y_i - q_{[\tau_{0,\mathbb{K}_0(X_i)-1},\tau_{0,\mathbb{K}_0(X_i)}),0}(X_i)\right\}\right]\right|.$$

The first line can be shown to be $o_p(n^{-1/2})$ using similar arguments in the proof of Part 1. The second line can be shown to be $o_p(n^{1/2})$ by noting that the difference between d_0 and $\widehat{d}$ is asymptotically negligible. This completes the proof.

B.10 Proof of Theorem 6

Before proving Theorem 6, it is worth mentioning that results in Lemma 1 and Lemma 4 do not rely on the assumption that $\theta_0(\cdot)$ is piecewise constant. These lemmas hold under the conditions in Theorem 6 as well. The proof is divided into two parts. In the first part, we derive the convergence rate of the integrated ℓ_2 loss for $\widehat{\theta}$. Then, we establish the convergence rate of the value under our I2DR.

Convergence rate of the integrated ℓ_2 loss: We first establish the upper error bound on the integrated ℓ_2 loss of $\widehat{\theta}(\cdot)$. Here, we consider a more general framework. Specifically, define

$$\mathrm{AE}_k(\theta_0) = \inf_{\substack{\mathcal{P}:|\mathcal{P}|\le k+1\\ (\theta_{\mathcal{I}})_{\mathcal{I}\in\mathcal{P}}\in\prod_{\mathcal{I}\in\mathcal{P}}\mathbb{R}^{p+1}}} \left\{\sup_{a\in[0,1]}\left\|\theta_0(a) - \sum_{\mathcal{I}\in\mathcal{P}}\theta_{\mathcal{I}}\mathbb{I}(a\in\mathcal{I})\right\|_2\right\}.$$

It describes how well $\theta_0(\cdot)$ can be approximated by a step function with at most k change points. Consider the following class of functions

$$\mathbb{B}^{\alpha_0} = \left\{\theta_0(\cdot) : \limsup_{k\to\infty} k^{\alpha_0}\mathrm{AE}_k(\theta_0) < \infty\right\},$$

for some $\alpha_0 > 0$. The parameter α_0 characterizes the speed of approximation as the number of change points increases. According to the discussion in Section 4.2.1, the class of Hölder continuous functions in Model II belongs to $\mathbb{B}^{\alpha_0}$. In the following, we show with probability at least $1 - O(n^{-2})$ that $\int_0^1 \|\widehat{\theta}(a) - \theta_0(a)\|_2^2 da \le \bar{c}\gamma_n^{2\alpha_0/(1+2\alpha_0)}$ for any $\theta_0(\cdot) \in \mathbb{B}^{\alpha_0}$.

Since $\theta_0(\cdot) \in \mathbb{B}^{\alpha_0}$, for some sequence $\{k_n\}_n$ that satisfies $k_n \to \infty$ as $n \to \infty$, there exists a piecewise constant function $\theta^*(\cdot)$ such that

$$\theta^*(a) = \sum_{\mathcal{I} \in \mathcal{P}^*} \theta^*_{\mathcal{I}} \mathbb{I}(a \in \mathcal{I}), \qquad \forall a \in [0,1],$$

for some partition $\mathcal{P}^*$ of $[0,1]$ with $|\mathcal{P}^*| \le k_n + 1$ and some $(\theta^*_{\mathcal{I}})_{\mathcal{I} \in \mathcal{P}^*} \in \prod_{\mathcal{I} \in \mathcal{P}^*} \mathbb{R}^{p+1}$, and

$$\sup_{\mathcal{I} \in \mathcal{P}^*} \sup_{a \in \mathcal{I}} \|\theta_0(a) - \theta^*_{\mathcal{I}}\|_2 \le \frac{c_4}{k_n^{\alpha_0}}, \tag{127}$$

for some constant $c_4 > 0$. Detailed choice of k_n will be given later. Combining (127) together with (27), we obtain that

$$\sup_{\mathcal{I} \in \mathcal{P}^*} \|\theta^*_{\mathcal{I}}\|_2 \le 2c_0, \tag{128}$$

for sufficiently large n.

Let $\{\tau^*_k\}_{k=1}^{|\mathcal{P}^*|-1}$ with $0 < \tau^*_1 < \tau^*_2 < \cdots < \tau^*_{|\mathcal{P}^*|-1} < 1$ be the locations of the change points in $J(\mathcal{P}^*)$. For $1 \le k \le |\mathcal{P}^*| - 1$, define τ^{**}_k such that $0 \le \tau^{**}_k - \tau^*_k < 1/m$ and $\tau^{**}_k \in \{1/m, 2/m, \ldots, 1\}$. Let k^*_n be the largest integer that satisfies $k^*_n \le |\mathcal{P}^*| - 1$ and $\tau^{**}_{k^*_n} < 1$. Apparently, $k^*_n \le k_n$. Set $\tau^*_0 = \tau^{**}_0 = 0$ and $\tau^*_{k^*_n+1} = \tau^{**}_{k^*_n+1} = 1$. Define a new partition $\mathcal{P}^{**} \in \mathcal{B}(m)$ and the set of vectors $(\theta^{**}_{\mathcal{I}})_{\mathcal{I} \in \mathcal{P}^{**}}$ as follows,

$$\begin{aligned}
&\mathcal{P}^{**} = \{[\tau^{**}_0, \tau^{**}_1), [\tau^{**}_1, \tau^{**}_2), \cdots, [\tau^{**}_{k^*_n}, \tau^{**}_{k^*_n+1}]\},\\
&\theta^{**}_{[\tau^{**}_k, \tau^{**}_{k+1})} = \theta^*_{[\tau^*_k, \tau^*_{k+1})}, \quad \forall k \in \{0, 1, \ldots, k^*_n - 1\} \text{ and } \theta^{**}_{[\tau^{**}_{k^*_n}, 1]} = \theta^*_{[\tau^*_{k^*_n}, \tau^*_{k^*_n+1})} \ \ (\text{or } \ \theta^*_{[\tau^*_{k^*_n}, 1]}).
\end{aligned}$$

Notice that it is possible that $[\tau^{**}_k, \tau^{**}_{k+1}) = \emptyset$ for some $k < k^*_n$.

Then, it follows from (128) that

$$\sup_{\mathcal{I} \in \mathcal{P}^{**}} \|\theta^{**}_{\mathcal{I}}\|_2 \le 2c_0, \tag{129}$$

Moreover, it follows from (27), (129) and the condition $m \asymp n$ that

$$\begin{aligned}
\sum_{\mathcal{I} \in \mathcal{P}^{**}} \int_{\mathcal{I}} \|\theta_0(a) - \theta^{**}_{\mathcal{I}}\|_2^2 da &\le \sum_{\mathcal{I} \in \mathcal{P}^*} \int_{\mathcal{I}} \|\theta_0(a) - \theta^*_{\mathcal{I}}\|_2^2 da + \frac{|\mathcal{P}^{**}|}{m} \sup_{a \in [0,1], \mathcal{I} \in \mathcal{P}^{**}} \|\theta_0(a) - \theta^{**}_{\mathcal{I}}\|_2^2\\
&\le c_4^2 k_n^{-2\alpha_0} + 9c_0^2(k_n+1)m^{-1} \le O(1)(k_n^{-2\alpha_0} + n^{-1}k_n), \qquad (130)
\end{aligned}$$

for sufficiently large n, where $O(1)$ denotes some positive constant.

Notice that

$$
\begin{aligned}
&\sum_{i=1}^n \sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2 = \sum_{i=1}^n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}_1})^2\\
= \quad & \sum_{i=1}^n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)(Y_i-\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2}+\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2}-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}_1})^2\\
= \quad & \sum_{i=1}^n \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_2)(Y_i-\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2})^2 + \underbrace{\sum_{i=1}^n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)(\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2}-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}_1})^2}_{\chi_5}\\
+ \quad & 2\underbrace{\sum_{i=1}^n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)(Y_i-\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2})\overline{X}_i^\top(\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1})}_{\chi_6}.
\end{aligned}
$$

By definition, we have

$$
\sum_{i=1}^n \sum_{\mathcal{I}\in\widehat{\mathcal{P}}} \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2 + n\gamma_n|\widehat{\mathcal{P}}| \le \sum_{i=1}^n \sum_{\mathcal{I}\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I})(Y_i-\overline{X}_i^\top\theta^{**}_{\mathcal{I}})^2 + n(k_n+1)\gamma_n.
$$

It follows that

$$
\chi_5 + 2\chi_6 + n\gamma_n|\widehat{\mathcal{P}}| \le n(k_n+1)\gamma_n. \tag{131}
$$

We now give a lower bound for χ_5. Similar to (55) and (56), we can show that the following event occurs with probability at least $1-O(n^{-2})$:

$$
\lambda_{\min}\left(\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\overline{X}_i\overline{X}_i^\top\right) \ge c_5 n|\mathcal{I}|, \tag{132}
$$

for some constant $c_5>0$, and any interval $\mathcal{I}\in\mathfrak{I}(m)$ that satisfies $|\mathcal{I}|\ge \bar{c}_0 n^{-1}\log n$ where the constant $\bar{c}_0$ is defined in Lemma 1. Under the event defined in (132), we obtain that

$$
\begin{aligned}
\chi_5 \ge & \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)\mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\ge\bar{c}_0n^{-1}\log n)(\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2}-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}_1})^2\\
& \ge c_5 n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\ge\bar{c}_0n^{-1}\log n)|\mathcal{I}_1\cap\mathcal{I}_2|\|\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}\|_2^2.
\end{aligned} \tag{133}
$$

In addition, under the events defined in (22) and Lemma 4, we have

$$
\sup_{\mathcal{I}\in\widehat{\mathcal{P}}}\|\widehat{\theta}_{\mathcal{I}}-\theta_{0,\mathcal{I}}\|_2 \le \sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \frac{c_0\sqrt{\log n}}{\sqrt{|\mathcal{I}|n}} \le \frac{c_0\sqrt{\log n}}{\sqrt{\bar{c}_3 n\gamma_n}} = o(1),
$$

since $\gamma_n \gg n^{-1}\log n$. In view of (27), we obtain that

$$
\sup_{\mathcal{I}\in\widehat{\mathcal{P}}}\|\widehat{\theta}_{\mathcal{I}}\|_2 \le 2c_0, \tag{134}
$$

for sufficiently large n. This together with (129) yields that

$$\begin{aligned}&\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}}\sum_{\mathcal{I}_2\in\mathcal{P}^{**}}\mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\le\bar{c}_0 n^{-1}\log n)|\mathcal{I}_1\cap\mathcal{I}_2|\|\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}\|_2^2\\ \le\ & (4c_0^2\bar{c}_0 n^{-1}\log n)\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}}\sum_{\mathcal{I}_2\in\mathcal{P}^{**}}\mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\le\bar{c}_0 n^{-1}\log n),\end{aligned}$$

with probability at least $1-O(n^{-2})$. Recall that $\mathcal{P}^{**}$ has at most k_n change points. The number of nonempty intervals $\mathcal{I}_1\cap\mathcal{I}_2$ is at most $k_n+1+|\widehat{\mathcal{P}}|$. Thus, we obtain that

$$\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}}\sum_{\mathcal{I}_2\in\mathcal{P}^{**}}\mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\le\bar{c}_0 n^{-1}\log n)|\mathcal{I}_1\cap\mathcal{I}_2|\|\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}\|_2^2\le(k_n+1+|\widehat{\mathcal{P}}|)(4c_0^2\bar{c}_0 n^{-1}\log n),$$

with probability at least $1-O(n^{-2})$. This together with (133) yields that

$$\chi_5\ge c_5 n\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}}\sum_{\mathcal{I}_2\in\mathcal{P}^{**}}|\mathcal{I}_1\cap\mathcal{I}_2|\|\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}\|_2^2-c_5(k_n+1+|\widehat{\mathcal{P}}|)(4c_0^2\bar{c}_0\log n),$$

with probability at least $1-O(n^{-2})$, or equivalently,

$$\chi_5\ge c_5 n\int_0^1\|\widehat{\theta}(a)-\theta^{**}(a)\|_2^2da-c_5(k_n+1+|\widehat{\mathcal{P}}|)(4c_0^2\bar{c}_0\log n), \tag{135}$$

with probability at least $1-O(n^{-2})$, where

$$\theta^{**}(a)=\sum_{\mathcal{I}\in\mathcal{P}^{**}}\theta^{**}_{\mathcal{I}}\mathbb{I}(a\in\mathcal{I}).$$

We now provide an upper bound for $|\chi_6|$. Notice that

$$\begin{aligned}\chi_6 &= \sum_{i=1}^n\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}}\sum_{\mathcal{I}_2\in\mathcal{P}^{**}}\mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)(Y_i-\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2})\overline{X}_i^\top(\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}) \qquad (136)\\ &= \underbrace{\sum_{i=1}^n\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}}\sum_{\mathcal{I}_2\in\mathcal{P}^{**}}\mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i^\top(\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1})}_{\chi_7}\\ &+ \underbrace{\sum_{i=1}^n\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}}\sum_{\mathcal{I}_2\in\mathcal{P}^{**}}\mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)\{\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2}\}\overline{X}_i^\top(\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1})}_{\chi_8}.\end{aligned}$$

It suffices to provide upper bounds for $|\chi_7|$ and $|\chi_8|$.

Under the event defined in (23), we obtain that

$$\begin{aligned}
&\left|\sum_{i=1}^n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)\mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\ge \bar{c}_0 n^{-1}\log n)\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i^\top(\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1})\right| \\
\le &\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \left\|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)\mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\ge \bar{c}_0 n^{-1}\log n)\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\right\|_2 \|\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}\|_2 \\
&\le \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \sqrt{c_0|\mathcal{I}_1\cap\mathcal{I}_2|n\log n}\|\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}\|_2 \le \frac{c_5 n}{16}\int_0^1 \|\widehat{\theta}(a)-\theta^{**}(a)\|_2^2 da \\
&+\frac{4c_0\log n}{c_5}\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} |\mathcal{I}_1\cap\mathcal{I}_2| \le \frac{c_5 n}{16}\int_0^1 \|\widehat{\theta}(a)-\theta^{**}(a)\|_2^2 da + 4c_0c_5^{-1}\log n,
\end{aligned}$$

where the third inequality is due to Cauchy-Schwarz inequality.

In addition, using similar arguments in (37) and (38), we have with probability at least $1-O(n^{-2})$ that, for any interval $\mathcal{I}\in\mathfrak{I}(m)$ that satisfies $|\mathcal{I}|\le \bar{c}_0 n^{-1}\log n$,

$$\left\|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I})\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\right\|_2 \le \bar{c}_5\log n, \tag{137}$$

for some constant $\bar{c}_5>0$. Since the number of nonempty intervals $\mathcal{I}_1\cap\mathcal{I}_2$ is at most $k_n+1+|\widehat{\mathcal{P}}|$, we obtain that

$$\begin{aligned}
&\left|\sum_{i=1}^n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)\mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\le \bar{c}_0 n^{-1}\log n)\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i^\top(\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1})\right| \\
\le &\sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \left\|\sum_{i=1}^n \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)\mathbb{I}(|\mathcal{I}_1\cap\mathcal{I}_2|\le \bar{c}_0 n^{-1}\log n)\{Y_i-\overline{X}_i^\top\theta_0(A_i)\}\overline{X}_i\right\|_2 \|\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}\|_2 \\
&\le (k_n+1+|\widehat{\mathcal{P}}|)(\bar{c}_5\log n)\sup_{\mathcal{I}_1\in\widehat{\mathcal{P}}}\sup_{\mathcal{I}_2\in\mathcal{P}^{**}}\|\theta^{**}_{\mathcal{I}_2}-\widehat{\theta}_{\mathcal{I}_1}\|_2 \le 4c_0(k_n+1+|\widehat{\mathcal{P}}|)(\bar{c}_5\log n),
\end{aligned}$$

with probability at least $1-O(n^{-2})$. It follows that

$$|\chi_7|\le \frac{c_5 n}{16}\int_0^1 \|\widehat{\theta}(a)-\theta^{**}(a)\|_2^2 da + 4c_0c_5^{-1}\log n + 4c_0(k_n+1+|\widehat{\mathcal{P}}|)(\bar{c}_5\log n)\tag{138}$$

with probability at least $1-O(n^{-2})$.

As for $|\chi_8|$, it follows from Cauchy-Schwarz inequality that

$$\begin{aligned}
|\chi_8| \quad \le \quad & \frac{1}{4}\sum_{i=1}^n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)(\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2}-\overline{X}_i^\top\widehat{\theta}_{\mathcal{I}_1})^2 \\
+ \quad & \underbrace{\sum_{i=1}^n \sum_{\mathcal{I}_1\in\widehat{\mathcal{P}}} \sum_{\mathcal{I}_2\in\mathcal{P}^{**}} \mathbb{I}(A_i\in\mathcal{I}_1\cap\mathcal{I}_2)\{\overline{X}_i^\top\theta_0(A_i)-\overline{X}_i^\top\theta^{**}_{\mathcal{I}_2}\}^2}_{\chi_9} = \frac{\chi_5}{4}+\chi_9.
\end{aligned}\tag{139}$$

Notice that

$$\chi_9 = \sum_{i=1}^{n} \sum_{\mathcal{I}\in\mathcal{P}^{**}} \mathbb{I}(A_i \in \mathcal{I})\{\overline{X}_i^\top \theta_0(A_i) - \overline{X}_i^\top \theta^{**}_{\mathcal{I}}\}^2$$

It follows from (48), (50), (130) and Cauchy-Schwarz inequality that

$$\begin{aligned}
\mathrm{E}(\chi_9) = n \sum_{\mathcal{I}\in\mathcal{P}^{**}} \mathrm{E}\mathbb{I}(A\in\mathcal{I})\{\overline{X}^\top \theta_0(A) - \overline{X}^\top \theta^{**}_{\mathcal{I}}\}^2 &\le n \sum_{\mathcal{I}\in\mathcal{P}^{**}} \mathrm{E}\|\overline{X}\|_2^2 \mathbb{I}(A\in\mathcal{I})|\theta_0(A) - \theta^{**}_{\mathcal{I}}|_2^2 \\
\le n \sum_{\mathcal{I}\in\mathcal{P}^{**}} \mathrm{E}(\mathrm{E}\|\overline{X}\|_2^2|A)\mathbb{I}(A\in\mathcal{I})|\theta_0(A) - \theta^{**}_{\mathcal{I}}|_2^2 &\le \omega^2 n \sum_{\mathcal{I}\in\mathcal{P}^{**}} \mathrm{E}\mathbb{I}(A\in\mathcal{I})\|\theta_0(A) - \theta^{**}_{\mathcal{I}}\|_2^2 \\
&\le C_0\omega^2 n \sum_{\mathcal{I}\in\mathcal{P}^{**}} \int_{\mathcal{I}} \|\theta_0(a) - \theta^{**}_{\mathcal{I}}\|_2^2 da \le O(1)(n\kappa_n^{-2\alpha_0} + \kappa_n),
\end{aligned}$$

where $O(1)$ denotes some positive constant. Using similar arguments in (68), we have for any integer $q \ge 2$ that

$$\begin{aligned}
\mathrm{E}\left(\sum_{\mathcal{I}\in\mathcal{P}^{**}} \mathbb{I}(A\in\mathcal{I})\{\overline{X}^\top\theta_0(A) - \overline{X}^\top\theta^{**}_{\mathcal{I}}\}^2\right)^q &\le \sum_{\mathcal{I}\in\mathcal{P}^{**}} \mathrm{E}\mathbb{I}(A\in\mathcal{I})\{\overline{X}^\top\theta_0(A) - \overline{X}^\top\theta^{**}_{\mathcal{I}}\}^{2q} \\
&\le q!c^q \sum_{\mathcal{I}\in\mathcal{P}^{**}} \int_{\mathcal{I}} \|\theta_0(a) - \theta^{**}_{\mathcal{I}}\|_2^2 da \le q!C^q(n\kappa_n^{-2\alpha_0} + \kappa_n),
\end{aligned}$$

for some constants $c, C > 0$. Using Bernstein's inequality, we have for any $t > 0$ that

$$\Pr(\chi_9 \ge \mathrm{E}\chi_9 + t) \le \exp\left(-\frac{1}{2}\frac{t^2}{tC + 2C^2(nk_n^{-2\alpha_0} + k_n)}\right).$$

We will require the sequence $\{k_n\}_n$ to satisfy $k_n \gg \log n$. Set $t_0 = 4C\sqrt{(nk_n^{-2\alpha_0} + k_n)\log n}$, we have

$$\frac{t_0^2}{t_0 C + 2C^2(nk_n^{-2\alpha_0} + k_n)} = \frac{8\sqrt{nk_n^{-2\alpha_0} + k_n}\log n}{2\sqrt{\log n} + \sqrt{nk_n^{-2\alpha_0} + k_n}} \ge 2\log n,$$

for sufficiently large n. Therefore, we obtain with probability at least $1 - O(n^{-2})$ that

$$\chi_9 \le O(1)(n\kappa_n^{-2\alpha_0} + \kappa_n) + 4C\sqrt{(nk_n^{-2\alpha_0} + k_n)\log n} = O(nk_n^{-2\alpha_0} + k_n).$$

This together with (136), (138) and (139) yields that

$$|\chi_6| \le \frac{c_5 n}{16}\int_0^1 \|\widehat{\theta}(a) - \theta^{**}(a)\|_2^2 da + c_6\{(k_n + |\widehat{\mathcal{P}}|)\log n + nk_n^{-2\alpha_0}\} + \frac{\chi_5}{4}.$$

with probability at least $1 - O(n^{-2})$, for some constant $c_6 > 0$ and sufficiently large n. In view of (131) and (135), we obtain with probability at least $1 - O(n^{-2})$ that $\chi_5 \le n\gamma_n(k_n + 1 - \widehat{\mathcal{P}}) + 2|\chi_6|$ and hence

$$\begin{aligned}
& \frac{3c_5 n}{8}\int_0^1 \|\widehat{\theta}(a) - \theta^{**}(a)\|_2^2 da \\
\le \quad & 2c_6\{(k_n + |\widehat{\mathcal{P}}|)\log n + nk_n^{-2\alpha_0}\} + n\gamma_n(k_n + 1 - \widehat{\mathcal{P}}).
\end{aligned} \tag{140}$$

Suppose $|\widehat{\mathcal{P}}| \ge 2k_n+1$. Under the event defined in (140), it follows from the condition $n\gamma_n \gg \log n$ that

$$n\gamma_n(k_n+1-|\widehat{\mathcal{P}}|)+2c_6(k_n+|\widehat{\mathcal{P}}|)\log n \le 3c_6|\widehat{\mathcal{P}}|\log n - 2^{-1}n\gamma_n|\widehat{\mathcal{P}}| \le 0,$$

for sufficiently large n, and hence

$$\frac{3c_5 n}{8}\int_0^1 \|\widehat{\theta}(a)-\theta^{**}(a)\|_2^2 da \le 2c_6 n k_n^{-2\alpha_0}. \tag{141}$$

Otherwise, suppose $|\widehat{\mathcal{P}}| \le 2k_n$. It follows from (140) that

$$\frac{3c_5 n}{8}\int_0^1 \|\widehat{\theta}(a)-\theta^{**}(a)\|_2^2 da \le 6c_6(k_n\log n + nk_n^{-2\alpha_0})+n\gamma_n k_n,$$

with probability at least $1-O(n^{-2})$. This together with (141) yields that

$$\int_0^1 \|\widehat{\theta}(a)-\theta^{**}(a)\|_2^2 da \le 16c_5^{-1}c_6 n^{-1}(k_n\log n + nk_n^{-2\alpha_0})+3\gamma_n k_n, \tag{142}$$

with probability at least $1-O(n^{-2})$.

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}\int_0^1 \|\widehat{\theta}(a)-\theta_0(a)\|_2^2 da &= \int_0^1 \|\widehat{\theta}(a)-\theta^{**}(a)+\theta^{**}(a)-\theta_0(a)\|_2^2 da\\ &\le 2\int_0^1 \|\widehat{\theta}(a)-\theta^{**}(a)\|_2^2 da + 2\int_0^1 \|\theta^{**}(a)-\theta_0(a)\|_2^2 da.\end{aligned}$$

In view of (130) and (142), we obtain that

$$\int_0^1 \|\widehat{\theta}(a)-\theta_0(a)\|_2^2 da = O(n^{-1}k_n\log n + k_n^{-2\alpha_0}+\gamma_n k_n) = O(k_n^{-2\alpha_0}+\gamma_n k_n), \tag{143}$$

with probability at least $1-O(n^{-2})$, where the last equality is due to the condition that $\gamma_n \gg n^{-1}\log n$. Set $k_n = \lfloor \gamma_n^{-(1+2\alpha_0)} \rfloor$ (the largest integer that is smaller than $\gamma_n^{-(1+2\alpha_0)}$), we obtain that

$$\int_0^1 \|\widehat{\theta}(a)-\theta_0(a)\|_2^2 da = O(\gamma_n^{2\alpha_0/(1+2\alpha_0)}),$$

with probability at least $1-O(n^{-2})$. The proof is hence completed.

Convergence rate of the value function: To derive the convergence rate of the value function under the proposed I2DR, we introduce the following lemma.

Lemma 8 *Assume conditions in Theorem 6 hold. Then for any interval $\mathcal{I} \in \mathfrak{I}(m)$ with $|\mathcal{I}| \ge \bar{c}_0 n^{-1}\log n$ and any interval $\mathcal{I}' \in \widehat{\mathcal{P}}$ with $\mathcal{I} \subseteq \mathcal{I}'$, we have with probability at least $1-O(n^{-2})$ that*

$$\|\theta_{0,\mathcal{I}}-\theta_{0,\mathcal{I}'}\|_2 \le 3\sqrt{c_5^{-1}\gamma_n|\mathcal{I}|^{-1}},$$

where the constant c_5 *is defined in* (132). ■

Recall by the definition of the value function that

$$V^{opt} - V^{\pi^*}(\widehat{d}) = \mathrm{E}\left(\sup_{a\in[0,1]} \overline{X}^\top \theta_0(a)\right) - \mathrm{E}\left(\int_{\widehat{d}(X)} \overline{X}^\top \theta_0(a)\pi^*(a;X,\widehat{d}(X))da\right)(144)$$

We begin by providing an upper bound for

$$\chi_{11} = \mathrm{E}\left(\sup_{a\in[0,1]} \overline{X}^\top \theta_0(a)\right) - \mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \overline{X}^\top \theta_{0,\mathcal{I}}\right).$$

It follows from (48) and Cauchy-Schwarz inequality that

$$\begin{aligned}\chi_{11} &= \mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}}\sup_{a\in\mathcal{I}} \overline{X}^\top \theta_0(a)\right) - \mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \overline{X}^\top \theta_{0,\mathcal{I}}\right) \qquad (145)\\ &\le \mathrm{E}\|\overline{X}\|_2 \sup_{\mathcal{I}\in\widehat{\mathcal{P}}}\sup_{a\in\mathcal{I}} \|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2\\ &\le \sqrt{\mathrm{E}\sum_{j=1}^{p+1}|\overline{X}^{(j)}|^2} \sup_{\mathcal{I}\in\widehat{\mathcal{P}}}\sup_{a\in\mathcal{I}} \|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2 \le (p+1)^{1/2}\omega \sup_{\mathcal{I}\in\widehat{\mathcal{P}}}\sup_{a\in\mathcal{I}} \|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2.\end{aligned}$$

Consider a sequence $\{d_n\}_n$ that satisfies $d_n \ge 0, \forall n$, $d_n \to 0$ as $n\to\infty$ and $d_n \gg n^{-1}\log n$. By the definition of Hölder continuous functions, we have for any $\mathcal{I}$ with $|\mathcal{I}| \le d_n$ that

$$\sup_{a_1,a_2\in\mathcal{I}} \|\theta_0(a_1)-\theta_0(a_2)\|_2 \le L \sup_{a_1,a_2\in\mathcal{I}} |a_1-a_2|^{\alpha_0} \le L d_n^{\alpha_0}.$$

It follows that

$$\begin{aligned}\sup_{a\in\mathcal{I}} \|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2 &\le \sup_{a\in\mathcal{I}} \left\|\theta_0(a) - \{\mathrm{E}\overline{X}\overline{X}^\top \mathbb{I}(A\in\mathcal{I})\}^{-1}\mathrm{E}\overline{X}\overline{X}^\top\theta_0(A)\mathbb{I}(A\in\mathcal{I})\right\|_2\\ &\le \sup_{a\in\mathcal{I}} \left\|\{\mathrm{E}\overline{X}\overline{X}^\top \mathbb{I}(A\in\mathcal{I})\}^{-1}\mathrm{E}\overline{X}\overline{X}^\top\mathbb{I}(A\in\mathcal{I})\{\theta_0(a)-\theta_0(A)\}\right\|_2 \qquad (146)\\ &\le \sup_{a\in\mathcal{I}} \left\|\{\mathrm{E}\overline{X}\overline{X}^\top \mathbb{I}(A\in\mathcal{I})\}^{-1}\mathrm{E}\overline{X}\overline{X}^\top\mathbb{I}(A\in\mathcal{I})\right\|_2 \sup_{a,a^*\in\mathcal{I}} \|\theta_0(a)-\theta_0(a^*)\|_2 \le L d_n^{\alpha_0},\end{aligned}$$

for any $\mathcal{I}$ that satisfies $|\mathcal{I}| \le d_n$.

Consider an interval $\mathcal{I} \in \mathfrak{I}(m)$ that satisfies $|\mathcal{I}| > d_n$. For any $a \in \mathcal{I}$, we can find an interval $\mathcal{I}' \subseteq \mathcal{I}$ with $d_n/2 \le |\mathcal{I}'| \le d_n$ and $\mathcal{I}' \in \mathfrak{I}(m)$ that covers a. Similar to (146), we have

$$\|\theta_{0,\mathcal{I}'} - \theta_0(a)\|_2 \le L d_n^{\alpha_0}. \qquad (147)$$

Since $d_n \gg n^{-1}\log n$, by Lemma 8, we have with probability at least $1 - O(n^{-2})$ that

$$\|\theta_{0,\mathcal{I}} - \theta_{0,\mathcal{I}'}\|_2 \le 3\sqrt{2c_5^{-1}\gamma_n d_n^{-1}}.$$

This together with (147) yields that

$$\sup_{a\in\mathcal{I}} \|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2 \le L d_n^{\alpha_0} + 3\sqrt{2c_5^{-1}\gamma_n d_n^{-1}},$$

for any $\mathcal{I}\in\mathfrak{I}(m)$ that satisfies $|\mathcal{I}|>d_n$. Combining this together with (146), we obtain that

$$\sup_{a\in\mathcal{I}} \|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2 \le L d_n^{\alpha_0} + 3\sqrt{2c_5^{-1}\gamma_n d_n^{-1}},$$

for any $\mathcal{I}\in\mathfrak{I}(m)$, with probability at least $1-O(n^{-2})$. Set $d_n \asymp \gamma_n^{(1+2\alpha_0)^{-1}}$, we have with probability at least $1-O(n^{-2})$ that

$$\sup_{a\in\mathcal{I}} \|\theta_0(a)-\theta_{0,\mathcal{I}}\|_2 \le O(1)\gamma_n^{\alpha_0/(1+2\alpha_0)},$$

for any $\mathcal{I}\in\mathfrak{I}(m)$, where $O(1)$ denotes some positive constant.

Therefore, we obtain with probability at least $1-O(n^{-2})$ that

$$\chi_{11} \le O(1)\gamma_n^{\alpha_0/(1+2\alpha_0)}, \tag{148}$$

where $O(1)$ denotes some positive constant. Similarly, we can show with probability at least $1-O(n^{-2})$ that

$$\chi_{12} = \mathrm{E}\overline{X}^\top \theta_{0,\widehat{d}(X)} - \mathrm{E}\left(\int_{\widehat{d}(X)} \overline{X}^\top \theta_0(a)\pi^*(a;X,\widehat{d}(X))da\right) \le O(1)\gamma_n^{\alpha_0/(1+2\alpha_0)},$$

where $O(1)$ denotes some positive constant. This together with (144) and (148) yields that,

$$V^{opt} - V^{\pi^*}(\widehat{d}) \le \mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \overline{X}^\top \theta_{0,\mathcal{I}}\right) - \mathrm{E}\overline{X}^\top \theta_{0,\widehat{d}(X)} + O(1)\gamma_n^{\alpha_0/(1+2\alpha_0)}, \tag{149}$$

with probability at least $1-O(n^{-2})$, where $O(1)$ denotes some positive constant.

Using similar arguments in (145), we can show that

$$\mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \overline{X}^\top \theta_{0,\mathcal{I}}\right) - \mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \overline{X}^\top \widehat{\theta}_{\mathcal{I}}\right) \le (p+1)^{1/2}\omega \sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \|\theta_{0,\mathcal{I}}-\widehat{\theta}_{\mathcal{I}}\|_2,$$

and

$$\mathrm{E}\overline{X}^\top \theta_{0,\widehat{d}(X)} - \mathrm{E}\overline{X}^\top \widehat{\theta}_{\widehat{d}(X)} \le (p+1)^{1/2}\omega \sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \|\theta_{0,\mathcal{I}}-\widehat{\theta}_{\mathcal{I}}\|_2.$$

Since $\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \overline{X}^\top \widehat{\theta}_{\mathcal{I}} = \overline{X}^\top \widehat{\theta}_{\widehat{d}(X)}$, we have

$$V^{opt} - V^{\pi^*}(\widehat{d}) \le 2(p+1)^{1/2}\omega \sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \|\theta_{0,\mathcal{I}}-\widehat{\theta}_{\mathcal{I}}\|_2 + O(1)\gamma_n^{\alpha_0/(1+2\alpha_0)}, \tag{150}$$

under the event defined in (149). It follows from Lemma 1 and 4 that

$$\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \|\theta_{0,\mathcal{I}} - \widehat{\theta}_{\mathcal{I}}\|_2 \le \frac{\sqrt{\log n}}{\sqrt{n\gamma_n}},$$

with probability at least $1 - O(n^{-2})$. This together with (150) yields that

$$V^{opt} - V^{\pi^*}(\widehat{d}) \le O(1)\left(\gamma_n^{\alpha_0/(1+2\alpha_0)} + \frac{\sqrt{\log n}}{\sqrt{n\gamma_n}}\right),$$

with probability at least $1-O(n^{-2})$, where $O(1)$ denotes some positive constant. Set $\gamma_n \asymp (n^{-1/2}\log^{1/2} n)^{(2\alpha_0+1)/(4\alpha_0+1)}$, we obtain that $V^{opt}-V^{\pi^*}(\widehat{d}) = O(n^{-\alpha_0/(1+4\alpha_0)}\log^{\alpha_0/(1+4\alpha_0)} n)$, with probability at least $1 - O(n^{-2})$. The proof is hence completed.

B.11 Proof of Lemma 8

For a given interval $\mathcal{I}' \in \widehat{\mathcal{P}}$, the set of intervals $\mathcal{I}$ considered in Lemma 8 can be classified into the following three categories.
Category 1: $\mathcal{I} = \mathcal{I}'$. Then it is immediate to see that $\|\theta_{0,\mathcal{I}} - \theta_{0,\mathcal{I}'}\|_2 = 0$ and the assertion automatically holds.
Category 2: There exists another interval $\mathcal{I}^* \in \mathfrak{I}(m)$ that satisfies $\mathcal{I}' = \mathcal{I}^* \cup \mathcal{I}$. Notice that the partition $\widehat{\mathcal{P}}^* = \widehat{\mathcal{P}} \cup \{\mathcal{I}^*\} \cup \mathcal{I} - \{\mathcal{I}'\}$ also belongs to $\mathcal{B}(m)$. By definition, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{I}_0\in\widehat{\mathcal{P}}^*} \mathbb{I}(A_i\in\mathcal{I}_0)(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}_0})^2 + \lambda_n|\mathcal{I}_0|\|\widehat{\theta}_{\mathcal{I}_0}\|_2^2 + \gamma_n|\widehat{\mathcal{P}}^*| \\
\ge\ &\frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{I}_0\in\widehat{\mathcal{P}}} \mathbb{I}(A_i\in\mathcal{I}_0)(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}_0})^2 + \lambda_n|\mathcal{I}_0|\|\widehat{\theta}_{\mathcal{I}_0}\|_2^2 + \gamma_n|\widehat{\mathcal{P}}|,
\end{aligned}$$

and hence

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2 + \lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I}^*)(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}^*})^2 + \lambda_n|\mathcal{I}^*|\|\widehat{\theta}_{\mathcal{I}^*}\|_2^2 \\
\ge \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I}')(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}'})^2 + \lambda_n|\mathcal{I}'|\|\widehat{\theta}_{\mathcal{I}'}\|_2^2 - \gamma_n.
\end{aligned}$$

It follows from the definition of $\widehat{\theta}_{\mathcal{I}^*}$ that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I}^*)(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}^*})^2 + \lambda_n|\mathcal{I}^*|\|\widehat{\theta}_{\mathcal{I}^*}\|_2^2 \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I}^*)(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}'})^2 + \lambda_n|\mathcal{I}^*|\|\widehat{\theta}_{\mathcal{I}'}\|_2^2.$$

Therefore, we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}})^2 + \lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2 \qquad (151)\\
\ge\ &\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i\in\mathcal{I})(Y_i - \overline{X}_i^\top\widehat{\theta}_{\mathcal{I}'})^2 + \lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}'}\|_2^2 - \gamma_n.
\end{aligned}$$

Category 3: There exist two intervals $\mathcal{I}^*, \mathcal{I}^{**} \in \mathfrak{I}(m)$ that satisfy $\mathcal{I}' = \mathcal{I}^* \cup \mathcal{I} \cup \mathcal{I}^{**}$. Using similar arguments in proving (151), we can show that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + \lambda_n |\mathcal{I}| \|\widehat{\theta}_{\mathcal{I}}\|_2^2 \geq \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2 + \lambda_n |\mathcal{I}| \|\widehat{\theta}_{\mathcal{I}'}\|_2^2 - 2\gamma_n.$$

Hence, regardless of whether $\mathcal{I}$ belongs to Category 2, or it belongs to Category 3, we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + \lambda_n |\mathcal{I}| \|\widehat{\theta}_{\mathcal{I}}\|_2^2 \\ \geq\ & \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2 + \lambda_n |\mathcal{I}| \|\widehat{\theta}_{\mathcal{I}'}\|_2^2 - 2\gamma_n \\ \geq\ & \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2 - 2\gamma_n. \end{aligned} \tag{152}$$

Notice that $|\mathcal{I}'| \geq |\mathcal{I}|$. Under the event defined in (22), we obtain that

$$\|\widehat{\theta}_{\mathcal{I}'} - \theta_{0,\mathcal{I}'}\|_2 \leq \frac{c_0\sqrt{\log n}}{\sqrt{|\mathcal{I}|n}}.$$

Similar to (33), we can show the following event occurs with probability at least $1 - O(n^{-2})$,

$$\left| \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})\overline{X}_i^\top (\widehat{\theta}_{\mathcal{I}'} - \theta_{0,\mathcal{I}'}) \right| \leq O(1) n^{-1} \log n, \tag{153}$$

where $O(1)$ denotes some positive constant. Similarly, using Cauchy-Schwarz inequality, we can show with probability at least $1 - O(n^{-2})$ that

$$\begin{aligned} & \left| \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}} - \overline{X}_i^\top \theta_{0,\mathcal{I}'})\overline{X}_i^\top (\widehat{\theta}_{\mathcal{I}'} - \theta_{0,\mathcal{I}'}) \right| \\ \leq\ & \frac{1}{4n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}} - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2 + \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})\{\overline{X}_i^\top (\widehat{\theta}_{\mathcal{I}'} - \theta_{0,\mathcal{I}'})\}^2 \\ \leq\ & \frac{1}{4n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}} - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2 + O(1) n^{-1} \log n, \end{aligned}$$

where $O(1)$ denotes some positive constant. This together with (153) yields

$$\begin{aligned} & \left| \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})\overline{X}_i^\top (\widehat{\theta}_{\mathcal{I}'} - \theta_{0,\mathcal{I}'}) \right| \\ \leq\ & \frac{1}{4n}\sum_{i=1}^n \mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}} - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2 + O(1) n^{-1} \log n, \end{aligned}$$

with probability at least $1 - O(n^{-2})$, where $O(1)$ denote some positive constant. Using similar arguments in proving (32), we can show the following event occurs with probability at least $1 - O(n^{-2})$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}'})^2 \geq \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2 \tag{154}$$
$$-\frac{1}{2n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}} - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2 - O(1)n^{-1}\log n,$$

where $O(1)$ denotes some positive constant.

In addition, it follows from the definition of $\widehat{\theta}_{\mathcal{I}}$ that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + \lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2 + \lambda_n|\mathcal{I}|\|\theta_{0,\mathcal{I}}\|_2^2.$$

By (27) and the condition that $\lambda_n = O(n^{-1}\log n)$, we obtain that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \widehat{\theta}_{\mathcal{I}})^2 + \lambda_n|\mathcal{I}|\|\widehat{\theta}_{\mathcal{I}}\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2 + O(1)n^{-1}\log n,$$

where $O(1)$ denotes some positive constant. This together with (152) and (154) yields

$$\begin{aligned} & \sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2 \qquad (155)\\ \geq \quad & \sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2 - 2n\gamma_n - O(1)\log n \\ & -\frac{1}{2}\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}} - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2, \end{aligned}$$

with probability at least $1 - O(n^{-2})$, where $O(1)$ denotes some positive constant.

Notice that

$$\begin{aligned} & \sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2 = \sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}} + \overline{X}_i^\top \theta_{0,\mathcal{I}} - \overline{X}_i^\top \theta_{0,\mathcal{I}'})^2 \\ = \quad & \sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2 + 2\underbrace{\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})(\overline{X}_i^\top \theta_{0,\mathcal{I}} - \overline{X}_i^\top \theta_{0,\mathcal{I}'})}_{\chi_{10}} \\ + \quad & \sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}'} - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2. \end{aligned}$$

Combining this with (155) yields that

$$\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}'} - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2 \leq 4n\gamma_n + O(1)\log n + 4|\chi_{10}|. \tag{156}$$

Under the event defined in (132), we obtain that

$$\sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})(\overline{X}_i^\top \theta_{0,\mathcal{I}'} - \overline{X}_i^\top \theta_{0,\mathcal{I}})^2 \geq c_5 n |\mathcal{I}| \|\theta_{0,\mathcal{I}'} - \theta_{0,\mathcal{I}}\|_2^2. \tag{157}$$

By the definition of $\theta_{0,\mathcal{I}}$, we have $\mathrm{E}\mathbb{I}(A \in \mathcal{I})(Y - \overline{X}^\top \theta_{0,\mathcal{I}})\overline{X} = 0$. Under the event defined in (23), it follows from Cauchy-Schwarz inequality that

$$\begin{aligned} |\chi_{10}| \leq \frac{2}{c_5 n |\mathcal{I}|} \left\| \sum_{i=1}^{n} \mathbb{I}(A_i \in \mathcal{I})(Y_i - \overline{X}_i^\top \theta_{0,\mathcal{I}})\overline{X}_i \right\|_2^2 + \frac{c_5 n}{8} |\mathcal{I}| \|\theta_{0,\mathcal{I}'} - \theta_{0,\mathcal{I}}\|_2^2 \\ \leq \frac{2c_0^2 \log n}{c_5} + \frac{c_5 n}{8} |\mathcal{I}| \|\theta_{0,\mathcal{I}'} - \theta_{0,\mathcal{I}}\|_2^2. \end{aligned}$$

This together with (156) and (157) yields that

$$|\mathcal{I}| \|\theta_{0,\mathcal{I}} - \theta_{0,\mathcal{I}'}\|_2^2 \leq \frac{8\gamma_n}{c_5} + O(1) n^{-1} \log n,$$

with probability at least $1 - O(n^{-2})$, where $O(1)$ denotes some positive constant. Since $\gamma_n \gg n^{-1} \log n$, for sufficiently large n, we obtain with probability at least $1 - O(n^{-2})$ that

$$|\mathcal{I}| \|\theta_{0,\mathcal{I}} - \theta_{0,\mathcal{I}'}\|_2^2 \leq 9 c_5^{-1} \gamma_n.$$

The proof is hence completed.

B.12 Proof of Theorem 7

The proof of Theorem 7 relies on the following result that is proven in Lemma E.4 of Cai et al. (2021)[3]: For any interval $\mathcal{I} \in \mathfrak{I}(m)$ with $|\mathcal{I}| \gg \gamma_n$ and any interval $\mathcal{I}' \in \widehat{\mathcal{P}}$ with $\mathcal{I} \subseteq \mathcal{I}'$, we have with probability approaching 1 (w.p.a.1.) that

$$\mathrm{E}|q_{\mathcal{I},0}(X) - q_{\mathcal{I}',0}(X)|^2 \leq \bar{C} |\mathcal{I}|^{-1} \gamma_n, \tag{158}$$

for some constant $\bar{C} > 0$.

The rest of the proof is divided into two parts. In the first part, we show assertion (i) in Theorem 7 holds. In the second part, we present the proof for assertion (ii) in Theorem 7. It is worth mentioning that results in Lemmas 5 and 7 do not rely on the assumption that $Q(\cdot)$ is piecewise function. These lemmas hold under the conditions in Theorem 7 as well.

Proof of Part 1: Consider a sequence $\{d_n\}_n$ such that $d_n \to 0$ and $d_n \gg \gamma_n$. We aim to show

$$\max_{\substack{a \in \mathcal{I}' \\ \mathcal{I}' \in \widehat{\mathcal{P}}}} \mathrm{E}[|Q(X,a) - \widehat{q}_{\mathcal{I}'}(X)|^2] = O_p\big(\gamma_n^{\frac{2\alpha_0}{2\alpha_0+1}}\big) + O_p\big((n\gamma_n)^{-\frac{2\beta}{2\beta+p}} \log^4 n\big),$$

where the expectation is taken with respect to the marginal distribution of X.

[3] See `https://openreview.net/attachment?id=rvKD3iqtBdk&name=supplementary_material`.

By Lemma 5, it suffices to show

$$\max_{\substack{a\in\mathcal{I}'\\ \mathcal{I}'\in\widehat{\mathcal{P}}}} \mathrm{E}|Q(X,a)-q_{\mathcal{I}',0}(X)|^2 = O_p\big(\gamma_n^{\frac{2\alpha_0}{2\alpha_0+1}}\big). \tag{159}$$

Suppose $|\mathcal{I}'|\ge d_n$. Then according to (158), we can find some $\mathcal{I}$ such that $|\mathcal{I}|=d_n$ and $a\in\mathcal{I}\subseteq\mathcal{I}'$,

$$\mathrm{E}|q_{\mathcal{I},0}(X)-q_{\mathcal{I}',0}(X)|^2 \le \bar{C}\frac{\gamma_n}{d_n}.$$

In addition, it follows from Hölder smoothness assumption that

$$\max_x \max_{\mathcal{I}} \max_{a\in\mathcal{I}} |Q(x,a)-q_{\mathcal{I}}(x)| \le \max_x \max_{\mathcal{I}} \max_{a_1,a_2\in\mathcal{I}} |Q(x,a_1)-Q(x,a_2)| = O(d_n^{\alpha_0}).$$

By setting d_n to proportional to $\gamma_n^{1/(1+\alpha_0)}$, it is immediate to see that (159) holds.

Next, suppose $|\mathcal{I}'|<\gamma_n^{1/(1+\alpha_0)}$. Then it follows from the Hölder smoothness condition that (159) is satisfied as well. This completes the proof for the result (i).

Proof of Part 2: This part follows the second part of the proof of Theorem 6. Recall by the definition of the value function that

$$V^{opt}-V^{\pi^*}(\widehat{d}) = \mathrm{E}\left(\sup_{a\in[0,1]} Q(X,a)\right) - \mathrm{E}\left(\int_{\widehat{d}(X)} Q(X,a)\pi^*(a;X,\widehat{d}(X))da\right), \tag{160}$$

where the expectation is taken with respect to the marginal distribution of X.

Using similar arguments in (145), it follows from the result in Part 1 that

$$\mathrm{E}\left(\sup_{a\in[0,1]} Q(X,a)\right) - \mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} q_{\mathcal{I},0}(X)\right) = O_p\big(\gamma_n^{-\frac{\alpha_0}{2\alpha_0+1}}\big). \tag{161}$$

Similarly, we can show that

$$\mathrm{E}Q(X,\widehat{d}(X)) - \mathrm{E}\left(\int_{\widehat{d}(X)} Q(X,a)\pi^*(a;X,\widehat{d}(X))da\right) = O_p\big(\gamma_n^{-\frac{\alpha_0}{2\alpha_0+1}}\big).$$

This together with (160) and (161) yields that,

$$V^{opt}-V^{\pi^*}(\widehat{d}) \le \mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} q_{\mathcal{I},0}(X)\right) - \mathrm{E}Q(X,\widehat{d}(X)) + O_p\big(\gamma_n^{-\frac{\alpha_0}{2\alpha_0+1}}\big). \tag{162}$$

Using similar arguments in (145), we can obtain that

$$\mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} q_{\mathcal{I},0}(X)\right) - \mathrm{E}\left(\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \widehat{q}_{\mathcal{I}}(X)\right) \le \bar{C}' \sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \sqrt{\mathrm{E}[|q_{\mathcal{I},0}(X)-\widehat{q}_{\mathcal{I}}(X)|^2]},$$

$$\mathrm{E}Q(X,\widehat{d}(X)) - \mathrm{E}\widehat{Q}(X,\widehat{d}(X)) \le \bar{C}' \sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \sqrt{\mathrm{E}[|q_{\mathcal{I},0}(X)-\widehat{q}_{\mathcal{I}}(X)|^2]},$$

for some constant $\bar{C}'>0$. Since $\sup_{\mathcal{I}\in\widehat{\mathcal{P}}} \widehat{q}_{\mathcal{I}}(X) = \widehat{Q}(X,\widehat{d}(X))$, it follows from Lemma 5 and (162) that

$$V^{opt}-V^{\pi^*}(\widehat{d}) = O_p\big(\gamma_n^{\frac{\alpha_0}{2\alpha_0+1}}\big) + O_p\big((n\gamma_n)^{-\frac{\beta}{2\beta+p}}\log^4 n\big).$$

The proof is completed by setting γ_n to be proportional to $n^{-1/\left(1+\frac{\beta(1+2\alpha_0)}{\alpha_0(p+2\beta)}\right)}$.