DEEP JUMP Q-EVALUATION FOR OFFLINE POLICY EVALUATION IN CONTINUOUS ACTION SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider off-policy evaluation (OPE) in continuous action domains, such as dynamic pricing and personalized dose finding. In OPE, one aims to learn the value under a new policy using historical data generated by a different behavior policy. Most existing works on OPE focus on discrete action domains. To handle continuous action space, we develop a brand-new deep jump Q-evaluation method for OPE. The key ingredient of our method lies in adaptively discretizing the action space using deep jump Q-learning. This allows us to apply existing OPE methods in discrete domains to handle continuous actions. Our method is further justified by theoretical results, synthetic and real datasets.

1 INTRODUCTION

Individualization proposes to leverage omni-channel data to meet individual needs. Individualized decision making plays a vital role in a wide variety of applications. Examples include customized pricing strategy in economics (Qiang & Bayati, 2016; Turvey, 2017), individualized treatment regime in medicine (Chakraborty, 2013; Collins & Varmus, 2015), personalized recommendation system in marketing (McInerney et al., 2018; Fong et al., 2018), etc. Prior to adopting any decision rule in practice, it is crucial to know the impact of implementing such a policy. In many applications, it is risky to run a policy online to estimate its value (see, e.g., Li et al., 2011). Off-policy evaluation (OPE) thus attracts a lot of attention by learning the policy value offline using logged historical data.

Despite the popularity of developing OPE methods with a finite set of actions (see e.g., Dudík et al., 2011; 2014; Swaminathan et al., 2017; Wang et al., 2017), less attention has been paid to continuous action domains, such as dynamic pricing (den Boer & Keskin, 2020) and personalized dose finding (Chen et al., 2016). Recently, a few OPE methods have been proposed to handle continuous actions (Kallus & Zhou, 2018; Sondhi et al., 2020; Colangelo & Lee, 2020). All these methods rely on the use of a kernel function to extend the inverse probability weighting (IPW) or doubly robust (DR) approaches developed in discrete action domains. They suffer from three limitations. First, the validity of these methods requires the conditional mean of the reward given the feature-action pair to be a smooth function over the action space. This assumption could be violated in applications such as dynamic pricing, where the expected demand for a product has jump discontinuities as a function of the charged price (den Boer & Keskin, 2020). Second, the value estimator could be sensitive to the choice of the bandwidth parameter in the kernel function. It remains challenging to select this hyperparameter. Kallus & Zhou (2018) proposed to tune this parameter by minimizing the mean squared error of the resulting value estimator. However, their method is extremely computationally intensive in moderate or high-dimensional feature space; see Section 5 for details. Third, these kernel-based methods typically use a single bandwidth parameter. This is sub-optimal in cases where the second-order derivative of the conditional mean function has an abrupt change in the action space; see the toy example in Section 3.1 for details.

To address these limitations, we develop a deep jump Q-evaluation (DJQE) method by integrating multi-scale change point detection (see e.g., Fryzlewicz, 2014), deep learning (LeCun et al., 2015) and OPE in discrete action domains. The key ingredient of our method lies in adaptively discretizing the action space using deep jump Q-learning. This allows us to apply IPW or DR methods to handle continuous actions. It is worth mentioning that our method does not require kernel bandwidth selection. Theoretically, we show it allows the conditional mean to be either a continuous or piecewise function of the action (see Theorems 1 and 2). Empirically, we show it outperforms existing state-of-the-art OPE methods in both synthetic and real datasets.

2 PRELIMINARIES

We first formulate the OPE problem. We next discuss the kernel-based OPE methods and multi-scale change point detection, since our proposal is closely related to them.

2.1 OFF-POLICY EVALUATION

The observed datasets can be summarized into $\{(X_i, A_i, Y_i)\}_{1\leq i\leq n}$ where $O_i = (X_i, A_i, Y_i)$ denotes the feature-action-reward triplet for the ith subject and n denotes the total sample size. We assume these data triplets are independent copies of some population variables (X, A, Y). Let $\mathcal{X}$ and $\mathcal{A}$ denote the feature and action space, respectively. We focus on the setting where $\mathcal{A}$ is one-dimensional, as in dynamic pricing and personalized dose finding. A deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$ determines the action to be assigned given the observed feature. We use b to denote the behavior policy that generates the observed data. Specifically, $b(\bullet|x)$ denotes the probability density or mass function of A given $X = x$, depending on whether A is continuous or not. Define the expected reward function conditional on the feature-action pair as

$$Q(x, a) = \mathbb{E}\{Y|X = x, A = a\}.$$

We refer to this function as the Q-function, to be consistent with the literature on developing individualized treatment regime (Murphy, 2003).

As standard in the OPE and the causal inference literature (see e.g., Chen et al., 2016), we assume the stable unit treatment value assumption (SUTVA), no unmeasured confounders assumption, and the positivity assumption are satisfied. These assumptions guarantee that a policy's value is estimable from the observed data. Specifically, for a given target policy π, its value can be represented by

$$V(\pi) = \mathbb{E}\{Q(X, \pi(X))\}.$$

The goal of the OPE is to learn the value $V(\pi)$ based on the observed data.

2.2 KERNEL-BASED OPE

For discrete action, Zhang et al. (2012) and Dudík et al. (2011) proposed a DR estimator of $V(\pi)$ by

$$\frac{1}{n}\sum_{i=1}^{n}\psi(O_i, \pi, \widehat{Q}, \widehat{b}) = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{Q}(X_i, \pi(X_i)) + \frac{\mathbb{I}(A_i = \pi(X_i))}{\widehat{b}(A_i|X_i)}\{Y_i - \widehat{Q}(X_i, A_i)\}\right], \tag{1}$$

where $\mathbb{I}$ denotes the indicator function, $\widehat{Q}$ and $\widehat{b}$ denote some estimators for the Q-function and the behavior policy. The second term $\widehat{b}^{-1}(A_i|X_i)\mathbb{I}(A_i = \pi(X_i))\{Y_i - \widehat{Q}(X_i, A_i)\}$ inside the bracket corresponds to an augmentation term. Its expectation equals zero when $\widehat{Q} = Q$. The purpose of adding this term is to offer additional protection against potential model misspecification of the Q-function. Such an estimator is doubly-robust in the sense that its consistency relies on either $\widehat{Q}$ or $\widehat{b}$ to be correctly specified. By setting $\widehat{Q} = 0$, equation 1 is reduced to the IPW estimator.

In continuous action domains, the indicator function $\mathbb{I}(A_i = \pi(X_i))$ equals zero almost surely. Consequently, naively applying equation 1 yields the plug-in estimator $\sum_{i=1}^{n}\widehat{Q}(X_i, \pi(X_i))/n$. To address this concern, the kernel-based OPE proposed to replace the indicator function in equation 1 with a kernel function $K\{(A_i - \pi(X_i))/h\}$ with some bandwidth parameter h, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_h(O_i, \pi, \widehat{Q}, \widehat{b}) = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{Q}(X_i, \pi(X_i)) + \frac{K\{(A_i - \pi(X_i))/h\}}{\widehat{b}(A_i|X_i)}\{Y_i - \widehat{Q}(X_i, A_i)\}\right].$$

The bandwidth h represents a trade-off. The variance of the resulting value estimator decays with h. Yet, its bias increases with h. More specifically, it follows from Theorem 1 of Kallus & Zhou (2018) that the leading term of the bias is equal to

$$h^2\frac{\int u^2K(u)du}{2}\mathbb{E}\left(\left.\frac{\partial^2 Q(X, a)}{\partial a^2}\right|_{a=\pi(X)}\right). \tag{2}$$

To ensure the term in 2 decays to zero as h goes to 0, it requires the expected second derivative of the Q-function exists, and thus $Q(x, a)$ needs to be a smooth function of a. However, as commented in the introduction, this assumption could be violated in applications such as dynamic pricing.

Table 1: The bias and the standard deviation (in parentheses) of the value estimator for $V^{(1)}$ and $V^{(2)}$, using DJQE and kernel-based methods. $n = 100$, $X, A \sim$ Unif$[0, 1]$, $Y|X, A \sim N\{Q(X, A), 1\}$. The target policy is given by $\pi(x) = x$.

Method	DJQE	Kernel (small h, $h = 0.8$)	Kernel (large h, $h = 2$)
$V^{(1)}(\pi)$	0.34 (0.13)	0.65 (0.17)	0.33 (0.09)
$V^{(2)}(\pi)$	0.11 (0.33)	0.03 (0.32)	1.21 (0.12)

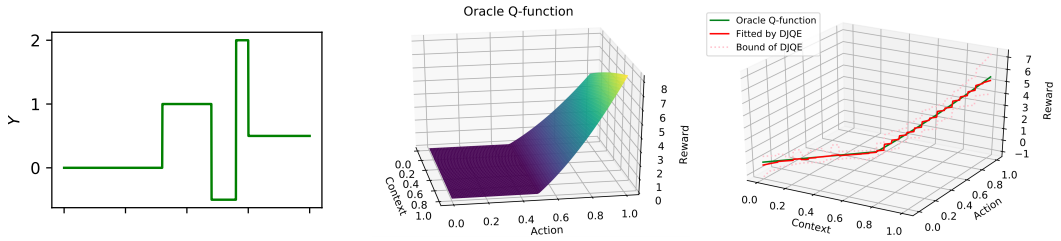

Figure 1: Left panel: example of piece-wise constant function in change point literature. Middle panel: the oracle Q-function on the feature-action space. Right Panel: the green curve presents the oracle Q-function $Q(x, \pi(x))$ under policy $\pi(x) = x$; and the red curve is the fitted mean value by DJQE and the pink dash line corresponds to the 95% confidence bound.

2.3 Multi-Scale Change Point Detection

The change point analysis considers an ordered sequence of data, $Y_{1:n} = \{Y_1, \cdots, Y_n\}$, with unknown change point locations, $\tau = \{\tau_1, \cdots, \tau_K\}$ for some unknown integer K. Here, τ_i is an integer between 1 and $n - 1$ inclusive, and satisfies $\tau_i < \tau_j$ for $i < j$. These change points split the data into $K + 1$ segments. Within each segments, the expected response is a constant function (see the left panel of Figure 1 for details). A number of methods have been proposed on estimating change points (see for example, Boysen et al., 2009; Killick et al., 2012; Frick et al., 2014; Fryzlewicz, 2014, and the references therein), by minimizing a penalized objective function:

$$\arg\min_{\tau} \left(\sum_{i=1}^{K+1} \left[\mathcal{C}\{Y_{(\tau_{i-1}+1):\tau_i}\} \right] + n\gamma K \right),$$

where $\mathcal{C}$ is a cost function that measures the goodness-of-the-fit of the constant function within each segment and $n\gamma K$ penalizes the number of change points. We remark that all the above cited works focused on models without features. Our proposal goes beyond these works in that we consider models with features and use deep neural networks (DNN) to capture the complex relationship between the response and features.

3 Deep Jump Q-Evaluation

In section 3.1, we use a toy example to demonstrate the limitation of kernel-based methods. We present the main idea of our algorithm in Section 3.2. Details are given in Section 3.3.

3.1 Toy Example

As discussed in the introduction, existing kernel-based OPE methods use a single bandwidth to construct the value estimator. Ideally, the bandwidth h in the kernel $K\{(A_i - \pi(X_i))/h\}$ shall vary with $\pi(X_i)$ to improve the accuracy of the value estimator. To elaborate this, consider the Q-function $Q(x, a) = 10 \max\{a^2 - 0.25, 0\} \log(x + 2)$ for any $x, a \in [0, 1]$. By definition, the Q-function is smooth over the entire feature-action space. However, it has different "patterns" when the action belongs to different intervals. Specifically, for $a \in [0, 0.5]$, $Q(x, a)$ is constant as a function of a. For $a \in (0.5, 1]$, $Q(x, a)$ depends quadratically in a. See the middle panel of Figure 1 for details.

Consider the target policy $\pi(x) = x$. We decompose the value $V(\pi)$ into $V^{(1)}(\pi) + V^{(2)}(\pi)$ where

$$V^{(1)}(\pi) = \mathbb{E}Q(X, \pi(X))\mathbb{I}(\pi(X) \leq 0.5) \text{ and } V^{(2)}(\pi) = \mathbb{E}Q(X, \pi(X))\mathbb{I}(\pi(X) > 0.5).$$

Similarly, denote the corresponding kernel-based value estimators by

$$\widehat{V}_h^{(1)}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\psi_h(O_i,\pi,\widehat{Q},\widehat{b})\mathbb{I}(\pi(X_i)\le 0.5) \text{ and } \widehat{V}_h^{(2)}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\psi_h(O_i,\pi,\widehat{Q},\widehat{b})\mathbb{I}(\pi(X_i)> 0.5).$$

Since $Q(x,a)$ is a constant function of $a \in [0,0.5]$, its second-order derivative $\partial^2 Q(x,a)/\partial a^2$ equals zero. In view of 2, when $\pi(x) \le 0.5$, the bias of $\widehat{V}_h^{(1)}(\pi)$ will be small even with a sufficiently large h. As such, a large h is preferred to reduce the variance of $\widehat{V}_h^{(1)}(\pi)$. When $\pi(x) > 0.5$, a small h is preferred to reduce the bias of $\widehat{V}_h^{(2)}(\pi)$. See Table 1 for details where we report the bias and standard deviation of $\widehat{V}_h^{(1)}(\pi)$ and $\widehat{V}_h^{(2)}(\pi)$ with two different bandwidths.

Due to the use of a single bandwidth, the kernel-based estimator suffers from either a large bias or a large variance. To overcome this limitation, we propose to adaptively discretize the action space into a union of disjoint intervals such that within each interval $\mathcal{I}$, the Q-function $\{Q(x,a) : a \in \mathcal{I}\}$ can be well-approximated by some function $Q_{\mathcal{I}}(x)$ that is constant in $a \in \mathcal{I}$. Based on the discretization, one can apply IPW or DR to evaluate the value. The advantage of adaptive discretization is illustrated in the right panel of Figure 1. When $a \le 0.5$, the Q-function is constant in a. It is likely that our procedure will not further split the interval $[0, 0.5]$. Consequently, the corresponding DR estimator for $V^{(1)}(\pi)$ will not suffer from large variance. When $a > 0.5$, our procedure will split $(0.5, 1]$ into a series of sub-intervals, approximating Q by a step function. This guarantees the resulting DR estimator for $V^{(2)}(\pi)$ will not suffer from large bias. Consequently, the proposed value estimator achieves a smaller mean squared error than kernel-based estimators. See Table 1 for details.

3.2 The Main Idea

For simplicity, we set the action space $\mathcal{A} = [0,1]$. From now on, we focus on a subset of intervals in $[0,1]$. By *interval* we always refer to those of the form $[c,d)$ for some $0 \le c < d < 1$, or $[c,1]$ for some $0 \le c < 1$, denoted as $\mathcal{I}$. A discretization $\mathcal{D}$ for $\mathcal{A}$ is defined as a collection of mutually disjoint intervals that covers $\mathcal{A}$. Let $|\mathcal{D}|$ denote the number of intervals in $\mathcal{D}$ and $|\mathcal{I}|$ denote the size of the interval $\mathcal{I}$. We aim to identify an "optimal" discretization $\widehat{\mathcal{D}}$ such that for each interval $\mathcal{I} \in \widehat{\mathcal{D}}$, $Q(x,a)$ is approximately a constant function of $a \in \mathcal{I}$. The number of intervals in $\widehat{\mathcal{D}}$ represents a trade-off. If $|\widehat{\mathcal{D}}|$ is too large, then $\widehat{\mathcal{D}}$ will contain many short intervals, the resulting IPW or DR estimator might suffer from large variance. Yet, a smaller value of $|\widehat{\mathcal{D}}|$ might result in a large bias. Our proposed method adaptively determines $\widehat{\mathcal{D}}$ and its size $|\widehat{\mathcal{D}}|$ as illustrated below.

To begin with, we cut the entire action space $\mathcal{A}$ into m initial intervals: $[0,1/m)$,$[1/m,2/m)$, $\ldots$, $[(m-1)/m,1]$. The number m shall be sufficiently large such that the Q-function can be well-approximated by a piecewise function on these intervals. In practice, we recommend to set the initial number of intervals m to be proportional to the sample size n. Note the set of these initial intervals is *not* the final partition $\widehat{\mathcal{D}}$ that we recommend, but only serve as the initial candidate intervals. We next adaptively combine some of these initial intervals to form the final partition $\widehat{\mathcal{D}}$. As shown in our numerical studies, the size of the final partition $|\widehat{\mathcal{D}}|$ is usually much less than m.

More specifically, denoted by $\mathcal{B}(m)$ as the set of discretizations $\mathcal{D}$ such that the end-points of each interval $\mathcal{I} \in \mathcal{D}$ lie on the grid $\{j/m : j = 0, 1, \cdots, m\}$. We associate to each partition $\mathcal{D} \in \mathcal{B}(m)$ a collection of functions $\{Q_{\mathcal{I}}\}_{\mathcal{I}\in\mathcal{D}}$. These functions depend only on features, not the action. They are used to produce a piecewise approximation of the Q-function such that $Q(a,\bullet) \approx \sum_{\mathcal{I}} \mathbb{I}(a \in \mathcal{I})Q_{\mathcal{I}}(\bullet)$. We model these $Q_{\mathcal{I}}$ using deep neural networks, to capture the complex dependence between the response and features. When the Q-function is well-approximated, we expect the least square loss $\sum_{\mathcal{I}\in\mathcal{D}}\left[\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})\big\{Y_i - Q_{\mathcal{I}}(X_i)\big\}^2\right]$, will be small.

Consequently, $\widehat{\mathcal{D}}$ can be estimated by solving

$$(\widehat{\mathcal{D}}, \{\widehat{q}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{D}}\}) = \underset{(\mathcal{D}\in\mathcal{B}(m),\{Q_{\mathcal{I}}\in\mathcal{Q}:\mathcal{I}\in\mathcal{D}\})}{\arg\min}\left(\sum_{\mathcal{I}\in\mathcal{D}}\left[\sum_{i=1}^{n}\mathbb{I}(A_i \in \mathcal{I})\big\{Y_i - Q_{\mathcal{I}}(X_i)\big\}^2\right] + n\gamma|\mathcal{D}|\right), \quad (3)$$

for some regularization parameter γ and DNN class $\mathcal{Q}$. Here, the penalty term $n\gamma|\mathcal{D}|$ in equation 3 controls the total number of intervals in $\widehat{\mathcal{D}}$, as in multi-scale change point detection. The tuning

parameter γ can be selected via cross-validation. We refer to this step as deep jump Q-learning. Details of this step are given in the next section. Given $\widehat{\mathcal{D}}$, one can apply IPW or DR (see equation 1) to derive the value estimates.

To further reduce the bias of the value estimator, we employ a data splitting and cross-fitting strategy, which is commonly used in statistics (Romano & DiCiccio, 2019). That is, we use different subsets of data samples to learn the discretization $\widehat{\mathcal{D}}$ and to construct the value estimator. A pseudocode summarizing our algorithm is given in Algorithm 1 in the appendix. We present the details below.

3.3 The Complete Algorithm

We present the details for DJQE in this section. It consists of three steps: data splitting, deep jump Q-learning, and cross-fitting.

Step 1: Data Splitting: We divide all n samples into L subsets of equal size, where $\mathbb{L}_\ell$ denotes the indices of samples in the ℓth subset for $\ell = 1, \cdots, L$. Let $\mathbb{L}_\ell^c = \{1, 2, \cdots, n\} - \mathbb{L}_\ell$ as the complement of $\mathbb{L}_\ell$.

Step 2: Deep Jump Q-Learning: For each $\ell = 1, \cdots, L$, we propose to apply deep jump Q-learning to compute a discretization $\widehat{\mathcal{D}}^{(\ell)}$ and $\{\widehat{q}_{\mathcal{I}}^{(\ell)} : \mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}\}$ by solving a version of equation 3 using the data subset in $\mathbb{L}_\ell^c$ only. We next present the computational details for solving this optimization. Our approach is motivated by the PELT method (Killick et al., 2012) in multi-scale change point detection.

Specifically, for any interval $\mathcal{I}$, define $\widehat{q}_{\mathcal{I}}^{(\ell)}$ as the minimizer of

$$\underset{Q_{\mathcal{I}} \in \mathcal{Q}}{\arg\min} \frac{1}{|\mathbb{L}_\ell^c|} \sum_{i \in \mathbb{L}_\ell^c} \mathbb{I}(A_i \in \mathcal{I}) \big\{ Q_{\mathcal{I}}(X_i) - Y_i \big\}^2, \tag{4}$$

where $|\mathbb{L}_\ell^c|$ denotes the number of samples in $\mathbb{L}_\ell^c$. Define the cost function $\mathcal{C}^{(\ell)}(\mathcal{I})$ as the minimum value of the objective function 4. In our implementation, we set $\mathcal{Q}$ as the class of multilayer perceptrons (MLPs, see Figure A in the appendix for an illustration) with L hidden layers and H nodes each hidden layer. The above optimization can be solved via the MLP regressor implementation of Pedregosa et al. (2011).

Computation of $\widehat{\mathcal{D}}^{(\ell)}$ relies on dynamic programming (Friedrich et al., 2008). For any integer $1 \leq v^* < m$, denote by $\mathcal{B}(m, v^*)$ the set consisting of all possible discretizations $\mathcal{D}_{v^*}$ of $[0, v^*/m)$. Set $\mathcal{B}(m, m) = \mathcal{B}(m)$, we define the Bellman function as

$$\text{Bell}(v^*) = \inf_{\mathcal{D}_{v^*} \in \mathcal{B}(m, v^*)} \left(\sum_{\mathcal{I} \in \mathcal{D}_{v^*}} \mathcal{C}^{(\ell)}(\mathcal{I}) + \gamma(|\mathcal{D}_{v^*}| - 1) \right), \text{ and } \text{Bell}(0) = -\gamma.$$

Our algorithm recursively updates the Bellman function based on the following formula,

$$\text{Bell}(v^*) = \min_{v \in \mathcal{R}_{v^*}} \left\{ \text{Bell}(v) + \mathcal{C}^{(\ell)}([v/m, v^*/m)) + \gamma \right\}, \quad \forall v^* \geq 1, \tag{5}$$

where $\mathcal{R}_{v^*}$ is the candidate change points list updated by

$$\{v \in \mathcal{R}_{v^*-1} \cup \{v^* - 1\} : \text{Bell}(v) + \mathcal{C}^{(\ell)}([v/m, (v^* - 1)/m)) \leq \text{Bell}(v^* - 1)\}, \tag{6}$$

during each iteration with $\mathcal{R}_0 = \{0\}$. The constraint listed in 6 is important as it facilitates the computation by discarding change points not relevant to obtain the final discretization, leading to a linear computational cost (Killick et al., 2012).

To solve equation 5, we search the optimal change point location v that minimizes $\text{Bell}(v^*)$. This requires to apply the MLP regressor to learn $\widehat{Q}^{(\ell)}_{[v/m, v^*/m)}$ and $\mathcal{C}^{(\ell)}([v/m, v^*/m))$ for each $v \in \mathcal{R}_{v^*}$. Let v^1 be the corresponding minimizer. We then define the change points list $\tau(v^*) = \{v^1, \tau(v^1)\}$. This procedure is iterated to compute $\text{Bell}(v^*)$ and $\tau(v^*)$ for $v^* = 1, \ldots, m$. The optimal partition $\widehat{\mathcal{D}}^{(\ell)}$ is determined by the values stored in τ (see Algorithm 1 in the appendix for details).

Step 3. Cross-Fitting: For each interval in the estimated optimal partition $\widehat{\mathcal{D}}^{(\ell)}$, we estimate the generalized propensity score function $\Pr(A \in \mathcal{I} | X = x)$ via the MLP regressor using the training dataset $\mathbb{L}_\ell^c$. Let $\widehat{b}^{(\ell)}(\mathcal{I}|x)$ denote the resulting estimate. The final estimated value for $V(\pi)$ is

constructed via cross-fitting, given by,

$$\widehat{V}(\pi) = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}\sum_{i\in\mathbb{L}_\ell}\left[\mathbb{I}(A_i\in\mathcal{I})\frac{\mathbb{I}\{\pi(X_i)\in\mathcal{I}\}}{\widehat{b}_{\mathcal{I}}^{(\ell)}(\mathcal{I}|X_i)}\{Y_i-\widehat{q}_{\mathcal{I}}^{(\ell)}(X_i)\}+\mathbb{I}(A_i\in\mathcal{I})\widehat{q}_{\mathcal{I}}^{(\ell)}(X_i)\right]. \quad (7)$$

Note the samples used to construct $\widehat{V}$ inside bracket are independent from those to estimate $\widehat{q}_{\mathcal{I}}^{(\ell)}$, $\widehat{b}_{\mathcal{I}}^{(\ell)}$ and $\widehat{\mathcal{D}}^{(\ell)}$. This helps remove the bias induced by overfitting in the estimation of $\widehat{q}_{\mathcal{I}}^{(\ell)}$, $\widehat{b}_{\mathcal{I}}^{(\ell)}$ and $\widehat{\mathcal{D}}^{(\ell)}$.

4 THEORY

We investigate the theoretical properties of the proposed estimator. For simplicity, let the support $\mathcal{X} = [0,1]^p$. We will show our estimator is consistent when the Q-function is either a piecewise function or a continuous function of a. Specifically, consider the following two model assumptions.

Model 1 (Piece-wise constant function). Suppose

$$Q(x,a) = \sum_{\mathcal{I}\in\mathcal{D}_0} q_{\mathcal{I}}(x)\mathbb{I}(a\in\mathcal{I}), \qquad \forall x\in\mathcal{X}, \quad \forall a\in\mathcal{A}, \quad (8)$$

for some partition $\mathcal{D}_0$ of $[0,1]$ and a collection of continuous functions $(q_{\mathcal{I}})_{\mathcal{I}\in\mathcal{D}_0}$.

Model 2 (Continuous function). Suppose Q is a continuous function of a and x.

We first consider the case where the value function takes the form of equation 8. We remark that kernel-based estimators will fail under this model assumption, as it requires the second-order derivative of the Q-function to exist. Without loss of generality, assume $q_{\mathcal{I}_1} \neq q_{\mathcal{I}_2}$ for any two adjacent intervals $\mathcal{I}_1, \mathcal{I}_2 \in \mathcal{D}_0$. This guarantees that the representation in equation 8 is unique. For any partition $\mathcal{D} = \{[0,\tau_1),[\tau_1,\tau_2),\cdots,[\tau_K,1]\}$, we use $J(\mathcal{D})$ to denote the set of change points $\{\tau_1,\cdots,\tau_K\}$. We impose the following conditions to establish our theories.

Assumption 1. The number of layers L and the number of nodes each hidden layer H diverge with n, in that $HL = O(n^\rho)$, for some constant $0<\rho<1/2$.

Assumption 2. Functions $\{\widehat{q}_{\mathcal{I}}^{(\ell)}(\bullet)\}_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}$ are uniformly bounded.

Assumption 1 is mild, as both H and L are parameters we specify. The part that $HL = O(n^\rho)$ ensures that the stochastic error resulting from the parameter estimation in the MLP is negligible. Assumption 2 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., 2018). These two assumptions guarantee the uniform consistency of the DNN estimator $\{\widehat{q}_{\mathcal{I}}\}_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}$. See Lemma 1 in the appendix for details.

Theorem 1 *Suppose Model 1, Assumptions 1 and 2 hold. Suppose m diverges to infinity with n. Then, there exists some constant γ_0 such that as long as $0<\gamma\le\gamma_0$, the following events occur with probability approaching 1 (w.p.a.1),*

(i) $|\widehat{\mathcal{D}}^{(\ell)}| = |\mathcal{D}_0|$; (ii) $\max_{\tau\in J(\mathcal{D}_0)}\min_{\hat\tau\in J(\widehat{\mathcal{D}}^{(\ell)})}|\hat\tau-\tau| = o_p(1)$. (iii) $\widehat{V}(\pi) = V(\pi)+o_p(1)$ for any policy π such that for any $\tau_0\in J(\mathcal{D})$, Pr$(\pi(X)\in[\tau_0-\epsilon,\tau_0+\epsilon]) \to 0$ as $\epsilon\to 0$.

Theorem 1 establishes the properties of our method under settings where the $Q(a,x)$ is piecewise function in a. Results in (i) imply that deep jump Q-learning correctly identifies the number of change points. Results in (ii) imply that for any change point τ in $\mathcal{D}_0$, there will be an estimated change point in $\widehat{\mathcal{D}}^{(\ell)}$ that converges to τ. Consequently, the change point locations can be consistently estimated. Combining (i) with (ii) yields the consistency of our estimated partition $\widehat{\mathcal{D}}^{(\ell)}$.

To ensure the consistency of the proposed value estimator, we require that the distribution of the random variable $\pi(X)$ induced by the target policy does not have point-masses at the change point locations. This condition is also mild. For instance, it automatically holds when $\pi(X)$ has a density function on $[0,1]$.

Theorem 2 *Suppose Model 2, Assumptions 1 and 2 hold. Suppose m diverges to infinity and γ decays to zero. Then we have*

(i) $\max_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}\sup_{a\in\mathcal{I}}\mathbb{E}|\widehat{q}_{\mathcal{I}}^{(\ell)}(X)-Q(X,a)|^2 = o(1)$; (ii) $\widehat{V}(\pi)-V(\pi) = o_p(1)$ for any π.

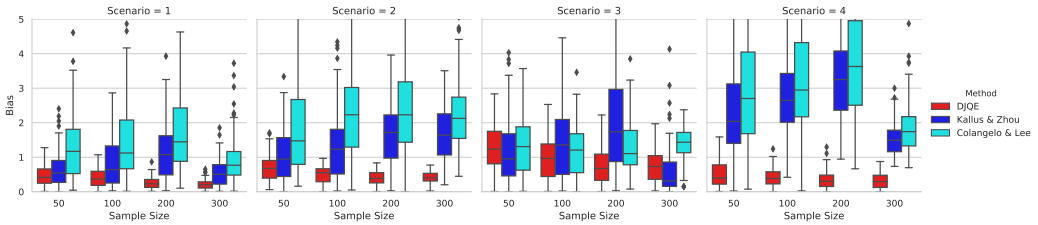

Figure 2: The box plot of the evaluated value function of the optimal policy under the proposed DJQE and two kernel-based methods for Scenario 1-4.

Theorem 2 establishes the properties of our method under settings where Q is continuous in a. Results in (i) imply that $\widehat{q}_{\mathcal{I}}^{(\ell)}(\bullet)$ can be used to uniformly approximate $Q(a, \bullet)$ for any $a \in \mathcal{I}$. The consistency of the value in (ii) thus follows.

5 Experiments

In this section, we investigate the finite sample performance of the proposed DJQE on the synthetic and real datasets, in comparison to two kernel-based methods. The computing infrastructure used is a virtual machine in the AWS Platform with 72 processor cores and 144GB memory. The average runtime for our proposed method on each setting is around 5 minutes.

5.1 Synthetic Data

Synthetic data are generated from the following model:

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

where $X = [X^{(1)}, X^{(2)}, \ldots, X^{(p)}]$. Consider the following different scenarios:
S1: $Q(x, a) = (1 + x^{(1)})\mathbb{I}(a < 0.35) + (x^{(1)} - x^{(2)})\mathbb{I}(0.35 \leq a < 0.65) + (1 - x^{(2)})\mathbb{I}(a \geq 0.65)$;
S2: $Q(x, a) = \mathbb{I}(a < 0.25) + \sin(2\pi x^{(1)})\mathbb{I}(0.25 \leq a < 0.5)\{0.5 - 8(x^{(1)} - 0.75)^2\}\mathbb{I}(0.5 \leq a < 0.75) + 0.5\mathbb{I}(a \geq 0.75)$;
S3 (toy): $Q(x, a) = 10 \max\{a^2 - 0.25, 0\} \log(x^{(1)} + 2)$;
S4: $Q(x, a) = 0.2(8 + 4x^{(1)} - 2x^{(2)} - 2x^{(3)}) - 2(1 + 0.5x^{(1)} + 0.5x^{(2)} - 2a)^2$.

The Q-function is a piecewise function of a under Scenarios 1 and 2, and is continuous under Scenarios 3 (toy example considered in Section 3.1) and 4. We set the target policy to be the optimal policy that achieves the highest possible mean reward. We list the oracle mean value under the optimal policy for each scenario in the first column of Table 3 in the appendix.

We apply the proposed DJQE and two kernel-based methods (Kallus & Zhou, 2018; Colangelo & Lee, 2020) to Scenario 1-4 with 20-dimensional covariates ($p = 20$) and $n \in \{50, 100, 200, 300\}$. For the DJQE, we set $m = n/10$ and select $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}n^{0.4}$ based on five-fold cross-validation. We find it extremely computationally intensive to compute the optimal bandwidth h^* using Kallus & Zhou (2018)'s method. For instance, it takes more than 48 hours to implement their method on a data with $n = 50$. Thus, as suggested in Kallus & Zhou (2018), we first compute h^* using data with sample size $n_0 = 50$. To accommodate data with different sample sizes n, we adjust h^* by setting $h^* = \{n_0/n\}^{0.2}$. To implement Colangelo & Lee (2020)'s estimator, we consider a list of bandwidths, given by $h = c\sigma_A n^{-0.2}$ with $c \in \{0.5, 0.75, 1.0, 1.5\}$ and σ_A is the sample standard deviation of the action. We then manually select the best bandwidth such that the resulting value estimator achieves the smallest mean squared error. The conditional mean value and generalized propensity score are fitted via the MLP regressor with 10 hidden layer and 10 neurons in each layer. The average estimated value and its standard deviation over 100 replicates are illustrated in Figure 2 for different methods, with detailed values reported in Table 3 in the appendix.

It can be seen from Figure 2 that the proposed DJQL is very efficient and outperforms all competing methods in almost all cases. We note that the proposed method performs reasonably well even when the sample size is small ($n = 50$). In contrast, two kernel-based methods fail to accurately estimate the value even in some cases when $n = 300$. Among the two kernel-based OPE approaches, we observe that the method developed by Kallus & Zhou (2018) performs better in general.

Table 2: The bias and the standard deviation (in parentheses) of the evaluated value function under the proposed DJQE and two kernel-based methods for the Warfarin data.

	n	100	300	500
Constant Policy	DJQE	0.121(0.167)	0.104(0.136)	0.131(0.221)
$V = -0.343$	Kallus & Zhou (2018)	1.998(0.316)	0.590(0.743)	0.786(0.871)
	Colangelo & Lee (2020)	6.660(9.620)	4.436(4.622)	5.706(5.608)
Linear Policy	DJQE	0.319(0.289)	0.335(0.501)	0.326(0.538)
$V = -0.325$	Kallus & Zhou (2018)	2.010(0.316)	0.590(0.741)	0.776(0.866)
	Colangelo & Lee (2020)	15.340(9.318)	5.909(5.962)	9.332(9.372)
Optimal Policy	DJQE	0.469(0.319)	0.468(0.574)	0.259(0.416)
$V = -0.278$	Kallus & Zhou (2018)	1.874(0.291)	0.531(0.659)	0.662(0.742)
	Colangelo & Lee (2020)	0.488(0.056)	0.225(0.713)	0.442(1.164)

5.2 Real Data: Personalized Dose Finding

Warfarin is commonly used for preventing thrombosis and thromboembolism. We use the dataset provided by the International Warfarin Pharmacogenetics (Consortium, 2009) for analysis. We choose $p = 81$ baseline covariates considered in Kallus & Zhou (2018). This yields a total of 3964 with complete records of baseline information. The response is defined as the absolute distance between the international normalized ratio (INR, a measurement of the time it takes for the blood to clot) after the treatment and the ideal value 2.5, i.e, $Y = -|\text{INR} - 2.5|$. We use the min-max normalization to convert the range of the dose level A into $[0, 1]$.

To compare among different methods, we calibrate the dataset to generate simulated outcomes. Specifically, we first estimate the Q-function via a MLP regressor with 10 hidden layer and 50 neurons in each layer using the whole dataset. We next use the fitted Q-function $\widehat{Q}(X, A)$ to simulate the data. Given a randomly sampled feature-action pair (a_j, x_j) from $\{(A_1, X_1), \cdots, (A_n, X_n)\}$, we set the reward r_j to $N\{\widehat{Q}(x_j, a_j), \widehat{\sigma}^2\}$, where $\widehat{\sigma}$ is the standard deviation of the fitted residual $\{Y_i - \widehat{Q}(X_i, A_i)\}_i$. Given the simulated data $\{(x_j, a_j, r_j) : 1 \leq j \leq n\}$, we are interested in evaluating the following three target policies:

(i) a **constant policy** $\pi^\star(X) \equiv 0$;

(ii) a **linear policy** $\pi^\star(X) = \frac{1}{p}\sum_{i=1}^{p=81} \frac{X^{(i)} - \inf(X^{(i)})}{\sup(X^{(i)}) - \inf(X^{(i)})}$;

(iii) an **optimal policy** $\pi^\star(X) \equiv \arg\max_{a \in [0,1]} \widehat{Q}(X, a)$.

The oracle values under these target policies are reported in Table 2.

We apply the DJQE on the calibrated Warfarin data, against two kernel-based methods. Due to the extremely intensive computation in Kallus & Zhou (2018), we directly apply the estimated optimal bandwidth h^* in their real data analysis, since they used the same dataset. Biases and standard deviations of the estimated values are reported in Table 2 over 20 replicates. A boxplot depicting these values is given in Figure 4 of the appendix.

It can be observed from Table 2 and Figure 4 that our proposed DJQE achieves much smaller mean squared error than the two kernel-based methods, when evaluating the constant or linear policy. When evaluating the optimal policy, Colangelo & Lee (2020)'s method performs the best with sample size $n = 100$. When $n = 300$, all three methods achieve similar mean squared error. When $n = 500$, our method performs the best.

6 Conclusion

We proposed a brand-new OPE algorithm in continuous action domains, via an adaptive discretization technique using deep jump Q-learning. In the current paper, we focus on settings with a single decision point. It would be practically interesting to extend our proposal to sequential decision making problems.

REFERENCES

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.

Bibhas Chakraborty. *Statistical methods for dynamic treatment regimes*. Springer, 2013.

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.

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.

Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

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.

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

Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 2018.

Christian Fong, Chad Hazlett, Kosuke Imai, et al. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

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.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. *arXiv preprint arXiv:1802.06037*, 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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 31–39, 2018.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:331–355, 2003.

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.

Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates. *Available at SSRN 2765257*, 2016.

J Romano and Cyrus DiCiccio. Multiple data splitting for testing. Technical report, Technical report, 2019.

Arjun Sondhi, David Arbour, and Drew Dimmery. Balanced off-policy evaluation in general action spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 2413–2423, 2020.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, pp. 3632–3642, 2017.

Ralph Turvey. *Optimal Pricing and Investment in Electricity Supply: An Esay in Applied Welfare Economics*. Routledge, 2017.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597, 2017.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68:1010–1018, 2012.

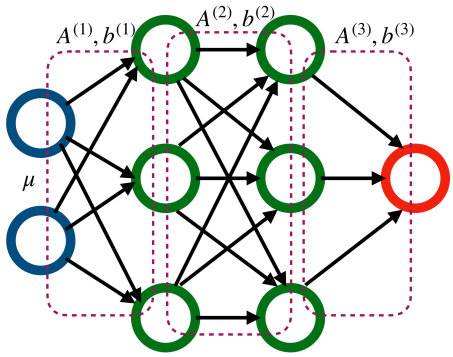

Figure 3: llustration of multilayer perceptron with two hidden layers and three nodes each hidden layer. Here μ is the input, $A^{(l)}$ and $b^{(l)}$ denote the corresponding parameters to produce the linear transformation for the $(l-1)$th layer.

A More on the implementation

We summarize our algorithm in Algorithm 1.

Global: data $\{(X_i, A_i, Y_i)\}_{1\le i\le n}$; number of initial intervals m; penalty term γ; target policy π.
Local: an upper triangular matrix of cost $\mathcal{C} \in \mathbb{R}^{m(m+1)/2}$; Bellman function Bell $\in \mathbb{R}^m$; partitions $\widehat{\mathcal{D}}$;
DNN functions $\{\widehat{q}_{\mathcal{I}}, \widehat{b}_{\mathcal{I}} : \mathcal{I} \in \widehat{\mathcal{D}}\}$; a vector $\tau \in \mathbb{N}^m$; a set of candidate point lists $\mathcal{R}$.
Output: the value estimator for target policy $\widehat{V}(\pi)$.
I. Split all n samples into L subsets as $\{\mathbb{L}_1, \cdots, \mathbb{L}_L\}$; $\widehat{V}(\pi) \leftarrow 0$;
II. Initialize an even segment on the action space with m pieces:
$\{\mathcal{I}\} = \{[0, 1/m), [1/m, 2/m), \ldots, [(m-1)/m, 1]\}$;
III. For $\ell = 1, \cdots, L$:
1. Set the training dataset as $\mathbb{L}_\ell^c = \{1, 2, \cdots, n\} - \mathbb{L}_\ell$;
2. Bell$(0) \leftarrow -\gamma$; $\widehat{\mathcal{D}} = [0, 1]$; $\tau \leftarrow Null$; $\mathcal{R}(0) \leftarrow \{0\}$;
3. Collect cost function:
For $r = 1, \ldots, m$: For $l = 0, \ldots, (r-1)$:
(i). Let $\mathcal{I} = [l/m, r/m)$ if $r < m$ else $\mathcal{I} = [l/m, 1]$;
(ii). Fit a MLP regressor: $\widehat{q}_{\mathcal{I}}(\cdot) \leftarrow \mathbb{I}(i \in \mathbb{L}_\ell^c)\mathbb{I}(A_i \in \mathcal{I})Y_i \sim \mathbb{I}(A_i \in \mathcal{I})MLP(X_i)$;
(iii). Calculate the cost: $\mathcal{C}(\mathcal{I}) \leftarrow \sum_{i\in\mathbb{L}_\ell^c} \mathbb{I}(A_i \in \mathcal{I})\{\widehat{q}_{\mathcal{I}}(X_i) - Y_i\}^2$;
4. Apply the PELT method to get partitions: For $v^* = 1, \ldots, m$:
(i).Bell$(v^*) = \min_{v\in\mathcal{R}(v^*)}\{$Bell$(v) + \mathcal{C}([v/m, v^*/m)) + \gamma\}$;
(ii). $v^1 \leftarrow \arg\min_{v\in\mathcal{R}(v^*)}\{$Bell$(v) + \mathcal{C}([v/m, v^*/m)) + \gamma\}$;
(iii). $\tau(v^*) \leftarrow \{v^1, \tau(v^1)\}$;
(iv). $\mathcal{R}(v^*) \leftarrow \{v \in \mathcal{R}(v^*-1) \cup \{v^*-1\} :$ Bell$(v) + \mathcal{C}([v/m, (v^*-1)/m)) \le$ Bell$(v^*-1)\}$;
5. Construct the DR value estimator: $r \leftarrow m$; $l \leftarrow \tau[r]$; While $r > 0$:
(i) Let $\mathcal{I} = [l/m, r/m)$ if $r < m$ else $\mathcal{I} = [l/m, 1]$; $\widehat{\mathcal{D}} \leftarrow \widehat{\mathcal{D}} \cup \mathcal{I}$;
(ii) Recall fitted MLP: $\widehat{q}_{\mathcal{I}}(\cdot) \leftarrow \mathbb{I}(i \in \mathbb{L}_\ell^c)\mathbb{I}(A_i \in \mathcal{I})Y_i \sim \mathbb{I}(A_i \in \mathcal{I})MLP(X_i)$;
(iii) Fit propensity score: $\widehat{b}_{\mathcal{I}}(\cdot) \leftarrow \mathbb{I}(i \in \mathbb{L}_\ell^c)\mathbb{I}(A_i \in \mathcal{I}) \sim \mathbb{I}(A_i \in \mathcal{I})MLP(X_i)$;
(iv) $r \leftarrow l$; $l \leftarrow \tau(r)$;
6. Evaluation using testing dataset $\mathbb{L}_\ell$:
$\widehat{V}(\pi) += \sum_{\mathcal{I}\in\widehat{\mathcal{D}}} \left(\sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I}) \left[\frac{\mathbb{I}\{\pi(X_i)\in\mathcal{I}\}}{\widehat{b}_{\mathcal{I}}(\mathcal{I}|X_i)}\{Y_i - \widehat{q}_{\mathcal{I}}(X_i)\} + \widehat{q}_{\mathcal{I}}(X_i)\right]\right)$;
return $\widehat{V}(\pi)/n$.

Algorithm 1: Deep Jump Q-Evaluation

B ADDITIONAL EXPERIMENT RESULTS

Table 3: The bias and the standard deviation (in parentheses) of the evaluated value function of the optimal policy under the proposed DJQE and two kernel-based methods for Scenario 1 to 4.

	n	50	100	200	300
Scenario 1	DJQE	0.445(0.381)	0.398(0.391)	0.253(0.269)	0.209(0.210)
V = 1.33	Kallus & Zhou (2018)	0.656(0.787)	0.848(0.799)	1.163(0.884)	0.537(0.422)
	Colangelo & Lee (2020)	1.285(1.230)	1.473(1.304)	1.826(1.463)	0.934(0.730)
Scenario 2	DJQE	0.696(0.376)	0.502(0.311)	0.400(0.219)	0.411(0.168)
V = 1.00	Kallus & Zhou (2018)	1.061(1.124)	1.363(1.131)	1.679(1.032)	1.664(0.792)
	Colangelo & Lee (2020)	1.827(1.371)	2.292(1.458)	2.429(1.541)	2.264(1.062)
Scenario 3	DJQE	2.014(0.865)	1.410(0.987)	1.184(0.967)	1.267(0.933)
V = 4.86	Kallus & Zhou (2018)	2.196(2.369)	2.758(2.510)	3.573(2.862)	1.151(1.798)
	Colangelo & Lee (2020)	2.586(2.825)	3.172(3.027)	3.949(3.391)	1.367(2.110)
Scenario 4	DJQE	0.494(0.485)	0.412(0.426)	0.349(0.383)	0.321(0.315)
V = 1.60	Kallus & Zhou (2018)	2.192(1.210)	2.740(1.034)	3.354(1.324)	1.555(0.500)
	Colangelo & Lee (2020)	2.975(1.789)	3.282(1.525)	3.921(1.927)	1.853(0.751)

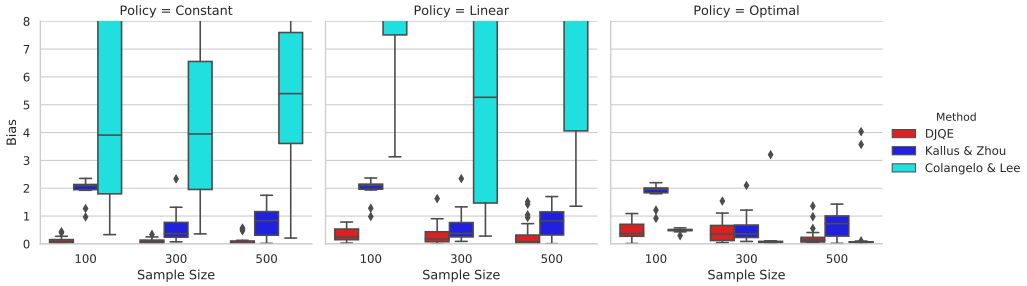

Figure 4: The box plot of the evaluated value function of the target policies under the proposed DJQE and two kernel-based methods for the Warfarin data.

C TECHNICAL PROOF

Throughout the proof, we use $c, C, c_0, \bar{c}$, etc., to denote some universal constants whose values are allowed to change from place to place. Let $O_i = \{X_i, Y_i\}$ denote the data summarized from the ith observation. For any interval $\mathcal{I}$, let $b(\mathcal{I}|x)$ denote the integral $\int_{a\in\mathcal{I}} b(a|x)da$.

Proofs of Theorems 1 and 2 rely on Lemmas 1 and 2. Specifically, Lemma 1 establishes the uniform convergence rate of $\widehat{q}_{\mathcal{I}}$ for any $\mathcal{I}$ whose length is no shorter than $o(\gamma)$ and belongs to the set of intervals:

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

To state this lemma, we first introduce some notations. For any such interval $\mathcal{I}$, define the function $q_{\mathcal{I}}(x) = \mathbb{E}(Y|A \in \mathcal{I}, X = x)$. It is immediate to see that the definition of $q_{\mathcal{I}}$ here is consistent with the one defined in equation 8 for any $\mathcal{I} \subseteq \mathcal{D}_0$.

Lemma 1 *The following holds when either the conditions in Theorem 1 or 2 hold:*

$$\max_{\mathcal{I}\in\mathfrak{I}(m), |\mathcal{I}|\ge c\gamma} \mathbb{E}[|q_{\mathcal{I}}(X) - \widehat{q}_{\mathcal{I}}^{(\ell)}(X)|^2 | \{O_i\}_{i\in\mathbb{L}_\ell}] = o_p(1),$$

for any positive constant $c > 0$.

Lemma 2 *When either the conditions in Theorem 1 or 2 hold,* $\min_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}|\mathcal{I}| \geq C\gamma$ *w.p.a.1 for some constant* $C > 0$.

We first present the proofs for these two lemmas. Next we present the proofs for Theorems 1 and 2.

C.1 Proof of Lemma 1

For any sufficiently small constant $\epsilon > 0$, it suffices to show

$$\max_{\mathcal{I}\in\mathfrak{I}(m),|\mathcal{I}|\geq c\gamma}\mathbb{E}[|q_{\mathcal{I}}(X)-\widehat{q}_{\mathcal{I}}^{(\ell)}(X)|^2|\{O_i\}_{i\in\mathbb{L}_\ell}]\leq C\epsilon^2,$$

w.p.a.1, for some constant $C > 1$ whose value will be specified later. The proof is divided into two parts. In the first part, we show there exist some constants $\{\theta^*_{\mathcal{I}}\}_{\mathcal{I}}$ such that

$$\max_{\mathcal{I}\in\mathfrak{I}(m)}\max_{x\in\mathcal{X}}|q_{\mathcal{I}}(x)-\text{MLP}(x;\theta^*_{\mathcal{I}})|\leq\epsilon. \tag{9}$$

In the second part, we show

$$\max_{\mathcal{I}\in\mathfrak{I}(m),|\mathcal{I}|\geq c\gamma}\mathbb{E}[|\text{MLP}(X;\theta^*_{\mathcal{I}})-\widehat{q}_{\mathcal{I}}^{(\ell)}(X)|^2|\{O_i\}_{i\in\mathbb{L}_\ell}]\leq(C-1)\epsilon^2. \tag{10}$$

The proof is hence completed.

Part 1. Under the step model assumption, for any $\mathcal{I}_0\in\mathcal{D}_0$, $q_{\mathcal{I}_0}(\bullet)$ is continuous. Since the support $\mathcal{X}$ is compact, $q_{\mathcal{I}_0}(\bullet)$ is uniformly continuous. Similarly, we can show b is uniformly continuous as well. For any $\mathcal{I}\in\mathfrak{I}_m$, we have

$$q_{\mathcal{I}}(x)=\frac{\int_{a\in\mathcal{I}}\sum_{\mathcal{I}_0\in\mathcal{D}_0}b(a|x)\mathbb{I}(a\in\mathcal{I}_0)q_{\mathcal{I}_0}(x)da}{b(\mathcal{I}|x)}.$$

It follows from the uniform continuity of $q_{\mathcal{I}_0}$, b and the positivity assumption that $q_{\mathcal{I}}(\bullet)$ is continuous for any $\mathcal{I}\in\mathfrak{I}(m)$.

Similarly, under the model assumption in Theorem 2, we can show $q_{\mathcal{I}}(\bullet)$ is continuous for any $\mathcal{I}\in\mathfrak{I}(m)$ as well. Consequently, when either the conditions in Theorem 1 or 2 hold, we obtain that $q_{\mathcal{I}}(\bullet)$ is continuous for any $\mathcal{I}\in\mathfrak{I}(m)$.

By Stone-Weierstrass theorem, there exists a multivariate polynomial function $q^*_{\mathcal{I}}$ such that the absolute value of the residual $q_{\mathcal{I}}-q^*_{\mathcal{I}}$ is uniformly bounded by $\epsilon/2$. By Theorem 1 of Yarotsky (2017), there exists a feedforward neural network with a bounded number of hidden units that uniformly approximates $q^*_{\mathcal{I}}$, with the approximation error uniformly bounded by $\epsilon/2$ in absolute value. By Lemma 1 of Farrell et al. (2018), such a feedforward network can be embedded into an MLP with a bounded number of hidden units. Since we allow H and L to diverge, such an MLP can be further embedded into an MLP with L layers and the widths of all layers being proportional to H. This yields equation 9. The proof for Part 1 is thus completed.

Part 2. We aim to show equation 10 holds. Under the boundedness assumption on Y, $\{q_{\mathcal{I}}(\bullet)\}_{\mathcal{I}\in\mathfrak{I}(m)}$ are uniformly bounded. We first observe that $\text{MLP}(\bullet,\theta^*_{\mathcal{I}})$ is a bounded function, by equation 9 and the fact that $q_{\mathcal{I}}(\bullet)$ is bounded.

Next, it follows from equation 9 that

$$\begin{aligned}&\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\text{MLP}(X;\theta^*_{\mathcal{I}})\}^2}{\Pr(A\in\mathcal{I})}=\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\mathbb{E}(Y|A\in\mathcal{I},X)\}^2}{\Pr(A\in\mathcal{I})}\\&+\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{\mathbb{E}(Y|A\in\mathcal{I},X)-\text{MLP}(X;\theta^*_{\mathcal{I}})\}^2}{\Pr(A\in\mathcal{I})}\leq\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\mathbb{E}(Y|A\in\mathcal{I},X)\}^2}{\Pr(A\in\mathcal{I})}+\epsilon^2.\end{aligned} \tag{11}$$

By definition,

$$\frac{\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\text{MLP}(X_i;\theta^*_{\mathcal{I}})\}^2}{\Pr(A\in\mathcal{I})}\geq\frac{\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\text{MLP}(X_i;\widehat{\theta}^{(\ell)}_{\mathcal{I}})\}^2}{\Pr(A\in\mathcal{I})}. \tag{12}$$

Suppose we can show

$$\sup_{\substack{\mathcal{I}\in\mathfrak{I}(m),|\mathcal{I}|\geq c\gamma\\ \theta_{\mathcal{I}}\in\Theta^*}}\left|\frac{\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\text{MLP}(X_i;\theta_{\mathcal{I}})\}^2}{|\mathbb{L}_\ell|\text{Pr}(A\in\mathcal{I})}-\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\text{MLP}(X;\theta_{\mathcal{I}})\}^2}{\text{Pr}(A\in\mathcal{I})}\right| = o_p(1), \tag{13}$$

where the set Θ^* consists of all $\theta_{\mathcal{I}}$ such that $\sup_x|\text{MLP}(x;\theta_{\mathcal{I}})|\leq M$ for some sufficiently large constant M such that both $\theta^*_{\mathcal{I}}$ and $\widehat{\theta}^{(\ell)}_{\mathcal{I}}$ satisfy this constraint. It thus follows from equation 11 and equation 12 that

$$\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\text{MLP}(X;\widehat{\theta}^{(\ell)}_{\mathcal{I}})\}^2}{\text{Pr}(A\in\mathcal{I})}\leq\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\mathbb{E}(Y|A\in\mathcal{I},X)\}^2}{\text{Pr}(A\in\mathcal{I})}+2\epsilon^2,$$

w.p.a.1. Similar to equation 11, we have

$$\begin{aligned}\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\text{MLP}(X;\widehat{\theta}^{(\ell)}_{\mathcal{I}})\}^2}{\text{Pr}(A\in\mathcal{I})}&=\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\mathbb{E}(Y|A\in\mathcal{I},X)\}^2}{\text{Pr}(A\in\mathcal{I})}\\&+\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{\mathbb{E}(Y|A\in\mathcal{I},X)-\text{MLP}(X;\widehat{\theta}^{(\ell)}_{\mathcal{I}})\}^2}{\text{Pr}(A\in\mathcal{I})}.\end{aligned}$$

It follows that

$$\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{\mathbb{E}(Y|A\in\mathcal{I},X)-\text{MLP}(X;\widehat{\theta}^{(\ell)}_{\mathcal{I}})\}^2}{\text{Pr}(A\in\mathcal{I})}=\frac{\mathbb{E}b(\mathcal{I}|X)\{\mathbb{E}(Y|A\in\mathcal{I},X)-\text{MLP}(X;\widehat{\theta}^{(\ell)}_{\mathcal{I}})\}^2}{\text{Pr}(A\in\mathcal{I})}\leq 2\epsilon^2.$$

Under the positivity assumption, $\sup_x b(\mathcal{I}|x)/\text{Pr}(A\in\mathcal{I})\geq 2(c-1)^{-1}$ for some constant $c>1$. This yields equation 10.

To complete the proof, it remains to show equation 13 holds.

Using similar arguments in Section A.2.2 of Farrell et al. (2018), for any $\mathcal{I}$, we have with probability at least $1-2\exp(-\bar{\gamma})$ that

$$\begin{aligned}&\left|\frac{\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\text{MLP}(X_i;\theta_{\mathcal{I}})\}^2}{|\mathbb{L}_\ell|\text{Pr}(A\in\mathcal{I})}-\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\text{MLP}(X;\theta_{\mathcal{I}})\}^2}{\text{Pr}(A\in\mathcal{I})}\right|\\&\leq\bar{c}\left\{\frac{M\bar{\gamma}}{n}+\sqrt{\frac{\bar{\gamma}}{n}}\sigma(\mathcal{I},\theta_{\mathcal{I}})+\sigma(\mathcal{I},\theta_{\mathcal{I}})\sqrt{\frac{H^2L^2}{n}}\left(\log\frac{M}{\sigma(\mathcal{I},\theta_{\mathcal{I}})}+\log n\right)\right\},\end{aligned}$$

for some constant $\bar{c}>0$, where $\sigma^2(\mathcal{I},\theta)$ corresponds to the variance of

$$\frac{1}{\text{Pr}(A\in\mathcal{I})}\mathbb{I}(A\in\mathcal{I})\{Y-\text{MLP}(X;\theta)\}^2.$$

Under the positivity assumption, we have $\sigma^2(\mathcal{I},\theta)\leq O(1)|\mathcal{I}|^{-1}$ where $O(1)$ denotes some positive constant. Consequently, for any $\mathcal{I}$ whose length is greater than $c\gamma^{-1}$, we have with probability at least $1-2\exp(-\bar{\gamma})$ that

$$\begin{aligned}&\left|\frac{\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\text{MLP}(X_i;\theta_{\mathcal{I}})\}^2}{|\mathbb{L}_\ell|\text{Pr}(A\in\mathcal{I})}-\frac{\mathbb{E}\mathbb{I}(A\in\mathcal{I})\{Y-\text{MLP}(X;\theta_{\mathcal{I}})\}^2}{\text{Pr}(A\in\mathcal{I})}\right|\\&\leq\bar{c}_0\left\{\frac{M\bar{\gamma}}{n}+\sqrt{\frac{\bar{\gamma}}{n\gamma}}+\sqrt{\frac{H^2L^2}{n\gamma}}(\log\gamma+\log n)\right\},\end{aligned}$$

for some constant $\bar{c}_0>0$.

Note that the number of elements in $\mathfrak{I}(m)$ is upper bounded by $(m+1)^2$. Since m is proportional to n, by setting $\bar{\gamma}=\bar{c}^*\log n$ for some constant $\bar{c}^*>0$, we have $1-2(m+1)^2\exp(-\bar{\gamma})\to 1$.

Consequently, it follows from Bonferroni's inequality that the following event occurs w.p.a.1 for any $\mathcal{I} \in \mathfrak{I}(m)$ such that $|\mathcal{I}| \geq c\gamma$,

$$\left| \frac{\sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \text{MLP}(X_i; \theta_{\mathcal{I}})\}^2}{|\mathbb{L}_\ell| \text{Pr}(A \in \mathcal{I})} - \frac{\mathbb{E}\mathbb{I}(A \in \mathcal{I})\{Y - \text{MLP}(X; \theta_{\mathcal{I}})\}^2}{\text{Pr}(A \in \mathcal{I})} \right| \leq O(1) \left\{ \frac{M \log n}{n} + \sqrt{\frac{\log n}{n\gamma}} + \sqrt{\frac{H^2 L^2}{n\gamma}} \log n \right\}, \tag{14}$$

where $O(1)$ denotes some positive constant. Under the given conditions, the RHS is $o(1)$. The proof is hence completed.

C.2 Proof of Lemma 2

Consider a given interval $\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}$. We can find some interval $\mathcal{I}' \in \mathfrak{I}(m) \cap \widehat{\mathcal{D}}^{(\ell)}$ that is adjacent to $\mathcal{I}$. Consequently, the interval $\mathcal{I} \cup \mathcal{I}'$ belongs to $\mathfrak{I}(m)$ as well. It follows that

$$\frac{1}{|\mathbb{L}_\ell|} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \widehat{q}_{\mathcal{I}}^{(\ell)}(X_i)\}^2 + \frac{1}{|\mathbb{L}_\ell|} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I}')\{Y_i - \widehat{q}_{\mathcal{I}'}^{(\ell)}(X_i)\}^2 \leq \frac{1}{|\mathbb{L}_\ell|} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')\{Y_i - \widehat{q}_{\mathcal{I} \cup \mathcal{I}'}^{(\ell)}(X_i)\}^2 - \gamma. \tag{15}$$

By definition, $\widehat{q}_{\mathcal{I} \cup \mathcal{I}'}^{(\ell)}$ minimizes the loss $\sum_{i \in \mathbb{L}_i} \mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')\{Y_i - \widehat{q}_{\mathcal{I} \cup \mathcal{I}'}^{(\ell)}(X_i)\}^2$. It follows that

$$\sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')\{Y_i - \widehat{q}_{\mathcal{I} \cup \mathcal{I}'}^{(\ell)}(X_i)\}^2 \leq \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I} \cup \mathcal{I}')\{Y_i - \widehat{q}_{\mathcal{I}'}^{(\ell)}(X_i)\}^2.$$

Combining this together with equation 15 yields that

$$\gamma \leq \frac{1}{|\mathbb{L}_\ell|} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \widehat{q}_{\mathcal{I}'}^{(\ell)}(X_i)\}^2. \tag{16}$$

As both Y and $\widehat{q}_{\mathcal{I}}^{(\ell)}$ are uniformly bounded, we have $\gamma \leq c|\mathbb{L}_\ell|^{-1} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})$ for some constant $c > 0$. By Bernstein's inequality, we have for any $t > 0$ such that

$$\frac{1}{|\mathbb{L}_\ell|} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I}) \leq \text{Pr}(A \in \mathcal{I}) + t, \tag{17}$$

with probability $1 - \exp(-nt^2/\bar{c}(|\mathcal{I}| + t))$ for some constant $\bar{c} > 0$.

Similar to equation 14, by setting $t_{\mathcal{I}} = \bar{c}_0 \max(\sqrt{|\mathcal{I}| n^{-1} \log n}, n^{-1} \log n)$ for some constant $\bar{c}_0 > 0$, we obtain w.p.a.1 that

$$\frac{1}{|\mathbb{L}_\ell|} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I}) \leq \text{Pr}(A \in \mathcal{I}) + t_{\mathcal{I}},$$

for any $\mathcal{I} \in \mathfrak{I}(m)$. Under the condition that b is continuous and $\mathcal{A} \times \mathcal{X}$ is compact, b is bounded. Consequently, the probability $\text{Pr}(A \in \mathcal{I})$ is proportional to the length of $\mathcal{I}$. In view of equation 17, for any $\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}$, it shall satisfy

$$\frac{\gamma}{\bar{c}_1} \leq |\mathcal{I}| + \frac{\log n}{n} + 2\sqrt{\frac{|\mathcal{I}| \log n}{n}} \leq 2|\mathcal{I}| + 2\frac{\log n}{n},$$

for some constant $\bar{c}_1 > 0$, w.p.a.1, where the last inequality follows from Cauchy-Schwarz inequality. As $\gamma \gg n^{-1} \log n$, we obtain $|\mathcal{I}| \geq \gamma/\bar{c}_2$ for any $\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}$ and some constant $\bar{c}_2 > 0$, w.p.a.1. The proof is hence completed.

C.3 PROOF OF THEOREM 1

We begin with an outline of the proof. The proof is divide into four steps. In the first step, we show

$$\Pr(|\widehat{\mathcal{D}}^{(\ell)}| \leq |\mathcal{D}_\ell|) \to 1. \tag{18}$$

In the second step, we show

$$\max_{\tau \in J(\mathcal{D}_0)} \min_{\widehat{\tau} \in J(\widehat{\mathcal{D}}^{(\ell)})} |\widehat{\tau} - \tau| < \delta_{\min}, \tag{19}$$

where $\delta_{\min} = \min_{\mathcal{I} \in \mathcal{D}_0} |\mathcal{I}|/3$. By the definition of $\delta_{\min}$, this implies that

$$\Pr(|\widehat{\mathcal{D}}^{(\ell)}| \geq |\mathcal{D}_\ell|) \to 1. \tag{20}$$

Combining equation 21 together with equation 18 proves (i) in Theorem 1. This proves (ii) in Theorem 1. In the third step, we show

$$\max_{\tau \in J(\mathcal{D}_0)} \min_{\widehat{\tau} \in J(\widehat{\mathcal{D}}^{(\ell)})} |\widehat{\tau} - \tau| = o_p(1). \tag{21}$$

In the last step, we show (iii) holds. The proof is thus completed.

We next detail the proof for each of the step.

Step 1. Assume $|\mathcal{D}_0| > 1$. Otherwise, equation 18 automatically holds. Consider the partition $\mathcal{D} = \{[0,1]\}$ which consists of a single interval and a zero q-function $q_{[0,1]}(x) = 0$ for any x. By definition, we have

$$\sum_{\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \widehat{q}_{\mathcal{I}}(X_i)\}^2 + |\mathbb{L}_\ell|\gamma|\widehat{\mathcal{D}}^{(\ell)}| \leq \sum_{i \in \mathbb{L}_\ell} Y_i^2 + |\mathbb{L}_\ell|\gamma.$$

Since Y is uniformly bounded and $\gamma = o(1)$, the right-hand-side (RHS) is $O(n)$. Consequently, we obtain

$$|\widehat{\mathcal{D}}^{(\ell)}| \leq c_0 \gamma^{-1}, \tag{22}$$

for some constant $c_0 > 0$.

Notice that

$$\begin{aligned}
\sum_{\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{Y_i - \widehat{q}_{\mathcal{I}}^{(\ell)}(X_i)\}^2 \geq & \underbrace{\sum_{\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{Y_i - q_{\mathcal{I}}(X_i)\}^2}_{\eta_1} \\
& + \underbrace{\sum_{\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{\widehat{q}_{\mathcal{I}}^{(\ell)}(X_i) - q_{\mathcal{I}}(X_i)\}^2}_{\eta_2} \\
& - \underbrace{2 \sum_{\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}} \left| \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{Y_i - q_{\mathcal{I}}(X_i)\}\{\widehat{q}_{\mathcal{I}}^{(\ell)}(X_i) - q_{\mathcal{I}}(X_i)\} \right|}_{\eta_3}.
\end{aligned} \tag{23}$$

We next show $\eta_2, \eta_3 = o_p(1)$.

Consider η_2 first. Under Lemma 2, we have w.p.a.1 that

$$\eta_2 = \sum_{\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}, |\mathcal{I}| \geq C\gamma} \sum_{i \in \mathbb{L}_\ell} \mathbb{I}(A_i \in \mathcal{I})\{\widehat{q}_{\mathcal{I}}^{(\ell)}(X_i) - q_{\mathcal{I}}(X_i)\}^2.$$

We decompose the RHS by

$$\begin{aligned}
& \sum_{\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}, |\mathcal{I}| \geq C\gamma} |\mathbb{L}_\ell| \mathbb{E}[\mathbb{I}(A \in \mathcal{I})\{\widehat{q}_{\mathcal{I}}^{(\ell)}(X) - q_{\mathcal{I}}(X)\}^2 | \{O_i\}_{i \in \mathbb{L}_i}] \\
+ & \sum_{\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}, |\mathcal{I}| \geq C\gamma} \sum_{i \in \mathbb{L}_\ell} [\mathbb{I}(A_i \in \mathcal{I})\{\widehat{q}_{\mathcal{I}}^{(\ell)}(X_i) - q_{\mathcal{I}}(X_i)\}^2 - \mathbb{E}\{\mathbb{I}(A \in \mathcal{I})\{\widehat{q}_{\mathcal{I}}^{(\ell)}(X) - q_{\mathcal{I}}(X)\}^2 | \{O_i\}_{i \in \mathbb{L}_i}\}].
\end{aligned}$$

The first line is $o_p(n)$, by Lemma 1. Using similar arguments in equation 14, we can show the second term is upper bounded by

$$\bar{c}|\mathbb{L}_\ell| \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)},|\mathcal{I}|\geq C\gamma} \left\{ \frac{M\log n}{n} + \sqrt{\frac{|\mathcal{I}|H^2L^2\log n}{n}} \right\},$$

for some constant $\bar{c} > 0$, w.p.a.1. In view of equation 22, the above expression can be further bounded by

$$\bar{c}n \left\{ \frac{c_0 M\log n}{\gamma n} + \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \sqrt{\frac{|\mathcal{I}|H^2L^2\log n}{n}} \right\} \leq O(1)n \left\{ \frac{\log n}{\gamma n} + \sqrt{\frac{H^2L^2\log n}{\gamma n}} \right\},$$

where $O(1)$ denotes some positive constant and the second inequality follows from the Cauchy-Schwarz inequality. This yields that $\eta_2 = o_p(n)$.

Using similar arguments, we can show that $\eta_3 = o_p(n)$. It follows that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i\in\mathcal{I})\{Y_i - \widehat{q}_{\mathcal{I}}^{(\ell)}(X_i)\}^2 \geq \eta_1 + o_p(n). \tag{24}$$

Notice that

$$\begin{aligned}
\eta_1 &= \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i\in\mathcal{I})\{Y_i - Q(A_i,X_i) + Q(A_i,X_i) - q_{\mathcal{I}}(X_i)\}^2 \\
&= \underbrace{\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i\in\mathcal{I})\{Y_i - Q(A_i,X_i)\}^2}_{\eta_4} + \underbrace{\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i\in\mathcal{I})\{Q(A_i,X_i) - q_{\mathcal{I}}(X_i)\}^2}_{\eta_5} \\
&+ \underbrace{2\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i\in\mathcal{I})\{Y_i - Q(A_i,X_i)\}\{Q(A_i,X_i) - q_{\mathcal{I}}(X_i)\}}_{\eta_6}.
\end{aligned}$$

Using similar arguments in bounding η_2 and η_3, we can show $\eta_6 = o_p(n)$ and that

$$\eta_5 = |\mathbb{L}_\ell| \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \int_{\mathcal{I}} \mathbb{E}b(a|X)|Q(X,a) - q_{\mathcal{I}}(X)|^2 da + o_p(n).$$

The first term on the RHS is greater than $c_0 n \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \int_{\mathcal{I}} \mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da$ for some constant $c_0 > 0$. It follows that

$$\eta_1 \geq \eta_4 + c_0|\mathbb{L}_\ell| \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \int_{\mathcal{I}} \mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da + o_p(n).$$

Note that under the step model assumption, η_4 can be rewritten as $\sum_{\mathcal{I}\in\mathcal{D}_0} \sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i\in\mathcal{I})\{Y_i - q_{\mathcal{I}}(X_i)\}^2$. This together with equation 24, yields that

$$\begin{aligned}
&\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i\in\mathcal{I})(Y_i - \widehat{q}_{\mathcal{I}}^{(\ell)}(X_i))^2 \\
\geq\ & \sum_{\mathcal{I}\in\mathcal{D}_0} \sum_{i\in\mathbb{L}_\ell} \mathbb{I}(A_i\in\mathcal{I})(Y_i - q_{\mathcal{I}}(X_i))^2 + c_0|\mathbb{L}_\ell| \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \int_{\mathcal{I}} \mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da + o_p(n),
\end{aligned}$$

For any integer k such that $1 \leq k \leq |\mathcal{D}_0|-1$, 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 $\mathcal{D}^*$ the oracle partition formed by the change point locations $\{\tau^*_{0,k}\}_{k=1}^{|\mathcal{D}_0|-1}$. Set $\tau^*_{0,0} = 0$, $\tau^*_{0,|\mathcal{D}_0|} = 1$ and $q^*_{[\tau^*_{0,k-1},\tau^*_{0,k})} = q_{[\tau_{0,k-1},\tau_{0,k})}$ for $1 \leq k \leq |\mathcal{D}_0| - 1$ and $q^*_{[\tau^*_{0,K-1},1]} = q_{[\tau_{0,K-1},1]}$. Let $\Delta_k = [\tau^*_{0,k-1},\tau^*_{0,k}) \cap [\tau_{0,k-1},\tau_{0,k})^c$ for

$1 \leq k \leq |\mathcal{D}_0| - 1$ and $\Delta_{|\mathcal{D}_0|} = [\tau^*_{0,|\mathcal{D}_0|-1}, 1] \cap [\tau_{0,|\mathcal{D}_0|-1}, 1]^c$. The length of each interval Δ_k is at most m^{-1}. It follows that

$$\begin{aligned}\sum_{\mathcal{I}\in\mathcal{D}^*}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i - q^*_{\mathcal{I}}(X_i)\}^2 - \sum_{\mathcal{I}\in\mathcal{P}_0}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i - q_{\mathcal{I}}(X_i)\}^2 \\ \leq \sum_{k=1}^{|\mathcal{D}_0|}\sum_{i\in\mathbb{L}_i}\mathbb{I}(A_i\in\Delta_k)\Big\{Y_i^2 + \sup_{\mathcal{I}\subseteq[0,1]}\sup_x q^2_{\mathcal{I}}(x)\Big\}.\end{aligned} \tag{25}$$

Using similar arguments in bounding the RHS of equation 16 in the proof of Lemma 2, we can show the RHS of equation 25 is $o_p(n)$, as m diverges to infinity. Combining this together with equation 24 yields

$$\begin{aligned}\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i - \widehat{q}^{(\ell)}_{\mathcal{I}}(X_i)\}^2 \geq \sum_{\mathcal{I}\in\mathcal{D}^*}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i - q^*_{\mathcal{I}}(X_i)\}^2 \\ + c_0|\mathbb{L}_\ell|\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}\int_{\mathcal{I}}\mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da + o_p(n).\end{aligned} \tag{26}$$

By definition, we have

$$\begin{aligned}&\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})(Y_i - \widehat{q}^{(\ell)}_{\mathcal{I}}(X_i))^2 + |\mathbb{L}_\ell|\gamma|\widehat{\mathcal{D}}^{(\ell)}| \\ \leq\ &\sum_{\mathcal{I}\in\mathcal{P}^*}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i - q^*_{\mathcal{I}}(X_i)\}^2 + |\mathbb{L}_\ell|\gamma|\mathcal{D}^*|.\end{aligned}$$

Since $|\mathcal{D}^*| = |\mathcal{D}_0|$, it follows from equation 26 that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}\int_{\mathcal{I}}\mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da \leq \gamma\{|\mathcal{D}_0| - |\widehat{\mathcal{D}}^{(\ell)}|\} + o_p(1). \tag{27}$$

Note that the left-hand-side (LHR) is non-negative. As $\gamma > 0$, we obtain that $|\widehat{\mathcal{D}}^{(\ell)}| \leq |\mathcal{D}_0|$, w.p.a.1. This completes the proof for the first step.

Step 2. It follows from equation 27 that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}\int_{\mathcal{I}}\mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da \leq \gamma\{|\mathcal{D}_0| - 1\} + o_p(1),$$

and hence

$$\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}}\int_{\mathcal{I}}\mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da \leq 2\gamma\{|\mathcal{D}_0| - 1\}, \tag{28}$$

w.p.a.1. We aim to show equation 19 holds under the event defined in equation 28. Otherwise, there exists some $\tau_0 \in J(\mathcal{P}_0)$ such that $|\hat{\tau} - \tau_0| \geq \delta_{\min}$, for all $\hat{\tau} \in J(\widehat{P})$. Under the event defined in equation 28, we obtain that

$$\int_{\tau_0-\delta_{\min}}^{\tau_0+\delta_{\min}}\mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da \leq 2\gamma(|\mathcal{D}_0| - 1), \tag{29}$$

w.p.a.1. On the other hand, since $Q(X,a)$ is a constant function on $[\tau_0 - \delta_{\min}, \tau_0)$ or $[\tau_0, \tau_0 + \delta_{\min})$, we have

$$\begin{aligned}&\int_{\tau_0-\delta_{\min}}^{\tau_0+\delta_{\min}}\mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da \\ \geq\ &\min_q\big(\delta_{\min}\mathbb{E}|q_{[\tau_0-\delta_{\min},\tau_0)}(X) - q(X)|^2 + \delta_{\min}\mathbb{E}|q_{[\tau_0,\tau_0+\delta_{\min})}(X) - q(X)|_2^2\big) \\ \geq\ &\frac{\delta_{\min}}{2}\mathbb{E}|q_{[\tau_0-\delta_{\min},\tau_0)}(X) - q_{[\tau_0,\tau_0+\delta_{\min})}(X)|^2 \geq \frac{\delta_{\min}\kappa_0}{2},\end{aligned}$$

where

$$\kappa_0 \equiv \min_{\substack{\mathcal{I}_1,\mathcal{I}_2\in\mathcal{P}_0 \\ \mathcal{I}_1 \text{ and } \mathcal{I}_2 \text{ are adjacent}}} \mathbb{E}|q_{\mathcal{I}_1}(X) - q_{\mathcal{I}_2}(X)|^2 2 > 0.$$

This apparently violates equation 29 where $\gamma \leq \delta_{\min}\kappa_0/(4|\mathcal{D}_0 - 4|)$. equation 19 thus holds w.p.a.1.

Step 3: Combining the results obtained in the first two steps yields that $|\mathcal{D}^{(\ell)}| = |\mathcal{D}_0|$, w.p.a.1. It follows from equation 27 that

$$\sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \int_{\mathcal{I}} \mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 da = o_p(1).$$

Using similar arguments in Step 2, we can show $\max_{\tau\in J(\mathcal{D}_0)} \min_{\widehat{\tau}\in J(\widehat{\mathcal{D}}^{(\ell)})} |\widehat{\tau} - \tau| < \epsilon$ w.p.a.1, for any sufficiently small $\epsilon > 0$. The proof is hence completed.

Step 4: We begin with some notations. For any π, we define a random policy $\pi_{\widehat{\mathcal{D}}^{(\ell)}}$ according to the partition $\widehat{\mathcal{D}}^{(\ell)}$ as follows:

$$\pi_{\widehat{\mathcal{D}}^{(\ell)}}(a|x) = \sum_{\mathcal{I}\subseteq\widehat{\mathcal{D}}^{(\ell)}} \mathbb{I}\{\pi(x)\in\mathcal{I}, a\in\mathcal{I}\}\frac{b(a|x)}{b(\mathcal{I}|x)}.$$

Note that $\int_0^1 \pi_{\widehat{\mathcal{D}}^{(\ell)}}(a|x)da = \sum_{\mathcal{I}\subseteq\widehat{\mathcal{D}}^{(\ell)}} \mathbb{I}\{\pi(x)\in\mathcal{I}\} = 1$ for any x. Consequently, $\pi_{\widehat{\mathcal{D}}^{(\ell)}}$ is a "valid" random policy.

The proposed value estimator $\widehat{V}^{(\ell)}$ is doubly-robust to $V(\pi_{\widehat{\mathcal{D}}^{(\ell)}})$. By Lemma 1, the estimated Q-function is consistent. Consequently, $\widehat{V}^{(\ell)}$ is consistent to $V(\pi_{\widehat{\mathcal{D}}^{(\ell)}})$. Since the propose estimator is a weighted average of $\widehat{V}^{(\ell)}$, it suffices to show $V(\pi_{\widehat{\mathcal{D}}^{(\ell)}})$ is consistent to $V(\pi)$. Note that

$$\begin{aligned} V(\pi_{\widehat{\mathcal{D}}^{(\ell)}}) &= \mathbb{E}\int_{[0,1]} Q(X,a) \sum_{\mathcal{I}\subseteq\widehat{\mathcal{D}}^{(\ell)}} \mathbb{I}\{\pi(X)\in\mathcal{I}, a\in\mathcal{I}\}\frac{b(a|X)}{b(\mathcal{I}|X)}da \\ &= \sum_{\mathcal{I}_0\in\mathcal{D}_0} \int_{\mathcal{I}_0} \mathbb{E}q_{\mathcal{I}_0}(X) \sum_{\mathcal{I}\subseteq\widehat{\mathcal{D}}^{(\ell)}} \mathbb{I}\{\pi(X)\in\mathcal{I}\}\frac{b(\mathcal{I}\cap\mathcal{I}_0|X)}{b(\mathcal{I}|X)}. \end{aligned}$$

Similarly, we can show

$$V(\pi) = \sum_{\mathcal{I}_0\in\mathcal{D}_0} \int_{\mathcal{I}_0} \mathbb{E}q_{\mathcal{I}_0}(X)\mathbb{I}\{\pi(X)\in\mathcal{I}_0\}. \tag{30}$$

For each $\mathcal{I}_0 \in \mathcal{D}_0$, there exists an interval $\mathcal{I} \in \widehat{\mathcal{D}}^{(\ell)}$ such that $|\mathcal{I}\cup\mathcal{I}_0|/|\mathcal{I}| \to 1$. We use $\widehat{\mathcal{I}}_0^{(\ell)}$ to denote this interval. Under the given conditions, we have

$$V(\pi_{\widehat{\mathcal{D}}^{(\ell)}}) = \sum_{\mathcal{I}_0\in\mathcal{D}_0} \int_{\mathcal{I}_0} \mathbb{E}q_{\mathcal{I}_0}(X)\mathbb{I}\{\pi(X)\in\widehat{\mathcal{I}}_0^{(\ell)}\} + o_p(1).$$

In view of equation 30, the value difference $|V(\pi_{\widehat{\mathcal{D}}^{(\ell)}}) - V(\pi)|$ can be upper bounded by

$$\sum_{\mathcal{I}_0\in\mathcal{D}_0} \Pr(\pi(X)\in\mathcal{I}_0 - \widehat{\mathcal{I}}_0^{(\ell)}) + \sum_{\mathcal{I}_0\in\mathcal{D}_0} \Pr(\pi(X)\in\widehat{\mathcal{I}}_0^{(\ell)} - \mathcal{I}_0) + o_p(1).$$

The first two terms decay to zero under the given conditions on $\pi(X)$. The proof is hence completed.

C.4 Proof of Theorem 2

We need the following lemma to prove Theorem 2.

Lemma 3 *Assume conditions in Theorem 2 hold. Then for any interval $\mathcal{I} \in \mathfrak{I}(m)$ with $|\mathcal{I}| \gg \gamma$ and any interval $\mathcal{I}' \in \widehat{\mathcal{D}}^{(\ell)}$ with $\mathcal{I} \subseteq \mathcal{I}'$, we have w.p.a.1 that*

$$\mathbb{E}|q_{\mathcal{I}}(X) - q_{\mathcal{I}'}(X)|^2 = o(1),$$

where the little-o term is uniform in $\mathcal{I}$ and $\mathcal{I}'$.

The rest of the proof is divided into three steps. In the first step, we show Assertion (i) in Theorem 2 holds. In the second step, we show Assertion (ii) in Theorem 2 holds. Finally, we present the proof for Lemma 3.

Step 1. Consider a sequence $\{d_n\}_n$ such that $d_n \to 0$ and $d_n \gg \gamma$. We aim to show $\inf_{\substack{a\in\mathcal{I}' \\ \mathcal{I}'\in\widehat{\mathcal{D}}^{(\ell)}}} \mathbb{E}[|Q(X,a) - \widehat{q}_{\mathcal{I}'}(X)|^2|\{O_i\}_{i\in\mathbb{L}_\ell}] = o_p(1)$. By Lemma 1, it suffices to show $\inf_{\substack{a\in\mathcal{I}' \\ \mathcal{I}'\in\widehat{\mathcal{D}}^{(\ell)}}} \mathbb{E}|Q(X,a) - q_{\mathcal{I}'}(X)|^2 = o(1)$.

Suppose $|\mathcal{I}'| \geq d_n$. Then according to Lemma 3, we can find some $\mathcal{I}$ such that $a \in \mathcal{I} \subseteq \mathcal{I}'$, $|\mathcal{I}| \to 0$, $\mathbb{E}|q_{\mathcal{I}}(X) - q_{\mathcal{I}'}(X)|^2 = o(1)$. Since $|\mathcal{I}| \to 0$ and $a \in \mathcal{I}$, it follows from the uniform continuity of the Q-function that $|q_{\mathcal{I}}(X) - Q(X,a)| \to 0$. The assertion thus follows.

Next, suppose $|\mathcal{I}'| < d_n$. Then $|\mathcal{I}'| \to 0$ as well. It follows from the uniform continuity of the Q-function that $\inf_{a\in\mathcal{I}'} |q_{\mathcal{I}'}(X) - Q(X,a)| \to 0$. The assertion thus follows. This completes the proof for the first step.

Step 2. Using similar arguments in Step 4 of the proof of Theorem 1, it suffices to show $V(\pi_{\widehat{\mathcal{D}}^{(\ell)}}) = V(\pi) + o_p(1)$. By definition,

$$\begin{aligned} V(\pi_{\widehat{\mathcal{D}}^{(\ell)}}) - V(\pi) &= \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \int_{\mathcal{I}} \mathbb{E}Q(X,a)\mathbb{I}(\pi(X)\in\mathcal{I})\frac{b(a|X)}{b(\mathcal{I}|X)}da - \mathbb{E}Q(\pi(X),X) \\ &= \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \int_{\mathcal{I}} \mathbb{E}\{Q(X,a) - Q(\pi(X),X)\}\mathbb{I}(\pi(X)\in\mathcal{I})\frac{b(a|X)}{b(\mathcal{I}|X)}da. \end{aligned}$$

It follows that

$$\begin{aligned} |V(\pi_{\widehat{\mathcal{D}}^{(\ell)}}) - V(\pi)| &\leq \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \inf_{a'\in\mathcal{I}} \int_{\mathcal{I}} \mathbb{E}|Q(a',X) - Q(\pi(X),X)|\mathbb{I}(\pi(X)\in\mathcal{I})\frac{b(a|X)}{b(\mathcal{I}|X)}da \\ &= \sum_{\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \inf_{a'\in\mathcal{I}} \mathbb{E}|Q(a',X) - Q(\pi(X),X)|\mathbb{I}(\pi(X)\in\mathcal{I}) \\ &\leq \inf_{a',a''\in\mathcal{I},\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \mathbb{E}|Q(a',X) - Q(a'',X)|. \end{aligned}$$

In Step 1 of the proof, we have shown that $\inf_{a\in\mathcal{I},\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \mathbb{E}|Q(X,a) - q_{\mathcal{I}}(X)|^2 = o(1)$. It follows that $\inf_{a,a'\in\mathcal{I},\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \mathbb{E}|Q(X,a) - Q(a',X)|^2 = o(1)$ and hence $\inf_{a',a''\in\mathcal{I},\mathcal{I}\in\widehat{\mathcal{D}}^{(\ell)}} \mathbb{E}|Q(a',X) - Q(a'',X)| = o(1)$, by Cauchy-Schwarz inequality. This completes the proof for the second step.

Step 3. For a given interval $\mathcal{I}' \in \widehat{\mathcal{D}}^{(\ell)}$, the set of intervals $\mathcal{I}$ considered in Lemma 3 can be classified into the following three categories.

Category 1: $\mathcal{I} = \mathcal{I}'$. It is immediate to see that $q_{\mathcal{I}} = q_{\mathcal{I}'}$ 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 $\widehat{\mathcal{D}}^{(\ell)*} = \widehat{\mathcal{D}}^{(\ell)} \cup \{\mathcal{I}^*\} \cup \mathcal{I} - \{\mathcal{I}'\}$ corresponds to another partition. By definition, we have

$$\begin{aligned} & \frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\sum_{\mathcal{I}_0\in\widehat{\mathcal{D}}^{(\ell)*}} \mathbb{I}(A_i\in\mathcal{I}_0)\{Y_i - \widehat{q}_{\mathcal{I}_0}(X_i)\}^2 + \gamma|\widehat{\mathcal{D}}^{(\ell)*}| \\ \geq\ & \frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\sum_{\mathcal{I}_0\in\widehat{\mathcal{D}}^{(\ell)}} \mathbb{I}(A_i\in\mathcal{I}_0)\{Y_i - \widehat{q}_{\mathcal{I}_0}(X_i)\}^2 + \gamma|\widehat{\mathcal{D}}^{(\ell)}|, \end{aligned}$$

and hence

$$\begin{aligned} & \frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i - \widehat{q}_{\mathcal{I}}(X_i)\}^2 + \frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I}^*)\{Y_i - \widehat{q}_{\mathcal{I}^*}(X_i)\}^2 \\ & \qquad \geq \frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I}')\{Y_i - \widehat{q}_{\mathcal{I}'}(X_i)\}^2 - \gamma. \end{aligned}$$

It follows from the definition of $\widehat{q}_{\mathcal{I}^*}$ that

$$\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I}^*)\{Y_i-\widehat{q}_{\mathcal{I}^*}(X_i)\}^2\leq\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I}^*)\{Y_i-\widehat{q}_{\mathcal{I}'}(X_i)\}^2.$$

Therefore, we obtain

$$\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2\geq\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}'}(X_i)\}^2-\gamma. \tag{31}$$

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 equation 31, we can show that

$$\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2\geq\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}'}(X_i)\}^2-2\gamma.$$

Hence, regardless of whether $\mathcal{I}$ belongs to Category 2, or it belongs to Category 3, we have

$$\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2\geq\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}'}(X_i)\}^2-2\gamma. \tag{32}$$

Using similar arguments in equation 14, we can show w.p.a.1 that

$$\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}}(X_i)\}^2=\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{Y-\widehat{q}_{\mathcal{I}}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]+o(\sqrt{\gamma|\mathcal{I}|}),$$
$$\frac{1}{|\mathbb{L}_\ell|}\sum_{i\in\mathbb{L}_\ell}\mathbb{I}(A_i\in\mathcal{I})\{Y_i-\widehat{q}_{\mathcal{I}'}(X_i)\}^2=\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{Y-\widehat{q}_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]+o(\sqrt{\gamma|\mathcal{I}|}),$$

where the little-o terms are uniform in $\mathcal{I}$ and $\mathcal{I}'$. Combining these together with equation 32 yields

$$\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{Y-\widehat{q}_{\mathcal{I}}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]\geq\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{Y-\widehat{q}_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]-2\gamma+o(\sqrt{\gamma|\mathcal{I}|}),$$

for any $\mathcal{I}$ and $\mathcal{I}'$, w.p.a.1. Note that $q_{\mathcal{I}}$ satisfies $\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{Y-q_{\mathcal{I}}(X)\}|X]=0$. We have

$$\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-\widehat{q}_{\mathcal{I}}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]\geq\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-\widehat{q}_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]-2\gamma+o(\sqrt{\gamma|\mathcal{I}|}).$$

Consider the first term on the RHS. Note that

$$\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-\widehat{q}_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]=\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-q_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]$$
$$+\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{\widehat{q}_{\mathcal{I}'}(X)-q_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]-2\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-q_{\mathcal{I}'}(X)\}\{\widehat{q}_{\mathcal{I}'}(X)-q_{\mathcal{I}'}(X)\}|\{O_i\}_{i\in\mathbb{L}_\ell}].$$

By Cauchy-Schwarz inequality, the last term on the RHS can be lower bounded by

$$-\frac{1}{2}\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-q_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]-2\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{\widehat{q}_{\mathcal{I}'}(X)-q_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}].$$

It follows that

$$\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-\widehat{q}_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]\geq\frac{1}{2}\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-q_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]$$
$$-3\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{\widehat{q}_{\mathcal{I}'}(X)-q_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}],$$

and hence

$$\frac{1}{2}\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-q_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]-2\gamma+o(\sqrt{\gamma|\mathcal{I}|})\leq\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-\widehat{q}_{\mathcal{I}}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]$$
$$+3\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}'}(X)-\widehat{q}_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}].$$

By Lemma 1 and the positivity assumption, the RHS is $o_p(|\mathcal{I}|)$. Note that the little-o terms are uniform in $\mathcal{I}$ and $\mathcal{I}'$. As $|\mathcal{I}|\gg\gamma$, we obtain that

$$\mathbb{E}[\mathbb{I}(A\in\mathcal{I})\{q_{\mathcal{I}}(X)-q_{\mathcal{I}'}(X)\}^2|\{O_i\}_{i\in\mathbb{L}_\ell}]=o_p(|\mathcal{I}|),$$

uniformly for any $\mathcal{I}$ and $\mathcal{I}'$, or equivalently,

$$\mathbb{E}\left[\frac{b(\mathcal{I}|X)}{|\mathcal{I}|}\{q_{\mathcal{I}}(X) - q_{\mathcal{I}'}(X)\}^2 | \{O_i\}_{i\in\mathbb{L}_\ell}\right] = o_p(1).$$

By the positivity assumption, we have

$$\mathbb{E}[\{q_{\mathcal{I}}(X) - q_{\mathcal{I}'}(X)\}^2 | \{O_i\}_{i\in\mathbb{L}_\ell}] = o_p(1),$$

uniformly for any $\mathcal{I}$ and $\mathcal{I}'$, or equivalently,

$$\mathbb{E}\{q_{\mathcal{I}}(X) - q_{\mathcal{I}'}(X)\}^2 = o_p(1).$$

This yields that

$$\mathbb{E}\{q_{\mathcal{I}}(X) - q_{\mathcal{I}'}(X)\}^2 = o(1),$$

uniformly for any $\mathcal{I}$ and $\mathcal{I}'$. The proof is hence completed.