QUANTILE-OPTIMAL POLICY LEARNING UNDER UNMEASURED CONFOUNDING

By

Zhongren Chen, Siyu Chen, Zhengling Qi,

Xiaohong Chen and Zhuoran Yang

June 2025

COWLES FOUNDATION DISCUSSION PAPER NO. 2469

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

arXiv:2506.07140v1 [stat.ML] 8 Jun 2025

Quantile-Optimal Policy Learning under Unmeasured Confounding

Zhongren Chen*[1], Siyu Chen*[1], Zhengling Qi†[2], Xiaohong Chen*[3], and Zhuoran Yang*[1]

[1] *Department of Statistics and Data Science, Yale University*
[2] *Department of Decision Sciences, George Washington University*
[3] *Cowles Foundation for Research in Economics, Yale University*

Abstract

We study quantile-optimal policy learning where the goal is to find a policy whose reward distribution has the largest α-quantile for some $\alpha \in (0,1)$. We focus on the offline setting whose generating process involves unobserved confounders. Such a problem suffers from three main challenges: (i) nonlinearity of the quantile objective as a functional of the reward distribution, (ii) unobserved confounding issue, and (iii) insufficient coverage of the offline dataset. To address these challenges, we propose a suite of causal-assisted policy learning methods that provably enjoy strong theoretical guarantees under mild conditions. In particular, to address (i) and (ii), using causal inference tools such as instrumental variables and negative controls, we propose to estimate the quantile objectives by solving nonlinear functional integral equations. Then we adopt a minimax estimation approach with nonparametric models to solve these integral equations, and propose to construct conservative policy estimates that address (iii). The final policy is the one that maximizes these pessimistic estimates. In addition, we propose a novel regularized policy learning method that is more amenable to computation. Finally, we prove that the policies learned by these methods are $\widetilde{\mathcal{O}}(n^{-1/2})$ quantile-optimal under a mild coverage assumption on the offline dataset. Here, $\widetilde{\mathcal{O}}(\cdot)$ omits poly-logarithmic factors. To the best of our knowledge, we propose the first sample-efficient policy learning algorithms for estimating the quantile-optimal policy when there exist unmeasured confounding.

1 Introduction

Offline reinforcement learning (RL) (Levine et al., 2020; Prudencio et al., 2023) aims to learn the optimal decision-making policies from pre-collected datasets, where the learner has no control over

*Emails: {zhongren.chen, siyu.chen.sc3226, xiaohong.chen, zhuoran.yang}@yale.edu
†Email: qizhengling@gwu.edu

the data collection process. This lack of control is further complicated by the fact that many real-world datasets suffer from incomplete information due to unmeasured confounders (Rubin, 1974)—factors influencing both the actions and rewards, yet not observed by the learner. These unobserved confounders naturally arise in the *partially observable* setting, where the behavior policy used to generate the actions is contingent on the state, while only noisy observations of the state are recorded in the data. For example, in healthcare, electronic health records (EHR) may lack key details about a patient's condition or treatment environment (Hernán and Robins, 2016), but decisions about treatment still need to be optimized based on the available data.

In many decision-making scenarios, rather than expected rewards, we are interested in optimizing more nuanced objectives involving quantiles of the reward distribution. For instance, to measure the efficacy of job training programs, it may be more meaningful to optimize for the median income increase, as the mean could be skewed by a few extreme cases. Moreover, quantile-based objectives naturally arise when some notion of fairness is of interest (Yang et al., 2019; Liu et al., 2022).

In this paper, we tackle the problem of quantile-optimal policy learning in the offline setting where unmeasured confounders play a critical role. Our goal is to find a policy that maximizes the α-th quantile of the reward distribution, conditioned on the context. Such a policy is learned based on a pre-collected offline dataset that involves unobserved confounders, which are hidden variables that simultaneously affect the rewards and actions stored in the data.

This problem is particularly challenging due to three core issues: (i) the quantile objective is a nonlinear functional of the reward distribution, making statistical learning and estimation more difficult; (ii) unmeasured confounders introduce bias into the learning problem, which can lead to misleading results if not properly accounted for; and (iii) the offline dataset often lacks full coverage, meaning that the distribution of the collected data might have insufficient overlap with that induced by some candidate policy, making it challenging to evaluate the performance of that policy.

To address these challenges, we propose a suite of novel, causal-assisted policy learning approaches. Our methods leverage powerful causal inference techniques, such as instrumental variables (IV) (Angrist et al., 1996; Baiocchi et al., 2014) and negative controls (NC) (Lipsitch et al., 2010; Tchetgen et al., 2020), to account for unmeasured confounding. Leveraging these causal inference tools, we reduce the quantile-based learning problem to the problem of solving **nonlinear functional integral equations**, which are then solved as a minimax estimation in nonparametric nonlinear conditional moment models. This enables us to address Challenges (i) and (ii). Furthermore, to handle Challenge (iii), we propose to adopt the pessimism principle (Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021; Buckman et al., 2020; Lu et al., 2022), which enables us to relax the requirement on data coverage. We introduce two algorithms based on the pessimism idea — a solution set constrained version and a regularized version. We note that pessimism penalizes candidate policies with large uncertainty and enables us to only focus on the policies that are

covered by the dataset. In particular, both our algorithms work as long as the offline data set has sufficient coverage over the optimal-policy induced distribution, regardless of whether the data set contains sufficient information about the suboptimal policies.

In this paper, we use quantile regret as the loss function. We establish that the proposed algorithms achieve sample-efficient $\widetilde{O}(n^{-1/2})$ suboptimality regret bounds, where n is the sample size. Here, $\widetilde{O}(\cdot)$ may include some logarithmic factors. Compared to existing works that study offline policy learning with confounded datasets aiming to maximize expected rewards, our analysis is more complicated due to the nonlinear nature of the quantile-based objective function. In particular, to get the suboptimality regret bounds, we leverage the local curvature of the nonlinear operator in the functional estimating equation. We examine the statistical error of the estimated quantile-based objective in terms of a local norm, defined using the first-order Taylor expansion of the nonlinear operator. To the best of our knowledge, we establish the first provably sample-efficient algorithms for quantile-optimal policy learning under unmeasured confounding and insufficient support.

Our contributions. Our contributions are threefold. First, leveraging causal tools such as IV and NC, we reduce the problem of quantile-optimal policy learning to a minimax estimation problem with nonlinear functional integral equations. To the best of our knowledge, we are the first to establish integral equations for the structural quantile function using NC. Second, to tackle the challenge of insufficient data coverage, we incorporate the principle of pessimism into our learning algorithm. Specifically, we introduce two versions of pessimistic algorithms, which are both proved to be sample efficient. Third, to analyze the pessimistic algorithms for quantile-optimal policy learning, we need to construct confidence sets for the nonlinear quantile-based minimax objective. To this end, we establish a novel statistical analysis framework that leverages the local curvature of the nonlinear operator in the functional estimating equation. Our analysis can be extended to general nonlinear functional integral equations, which may be of independent interest.

1.1 Related Literature

Our work belongs to the intersection of offline decision making, quantile policy learning, causal inference with unobserved confounders, and limited support.

Offline decision making There is already a large literature on offline reinforcement learning (Yin et al., 2022; Uehara et al., 2021; Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021) and offline contextual bandits (Li et al., 2012; Lee et al., 2021; Metevier et al., 2019). Within this literature, Cassel et al. (2023); Zhu and Tan (2020); Prashanth et al. (2020) study nonlinear objective functions in contextual bandits without confounding bias. Chen et al. (2023); Wang et al. (2021) investigate causal policy learning with confounding variables, but using linear objective

functions. Our work makes a new contribution to this literature by considering nonlinear objective functions with unobserved confounding and insufficient data coverage.

Quantile policy learning Quantile treatment effect has been extensively studied in Econometrics (Abadie et al., 2002; Chernozhukov and Hansen, 2005; Horowitz and Lee, 2007; Chernozhukov and Hansen, 2008; Chen and Pouzo, 2012; Gagliardini and Scaillet, 2012). Recently, a few papers have explored quantile policy learning. Wang et al. (2018) investigates quantile-optimal treatment regimes; Linn et al. (2017) proposes quantile regression to indirectly and approximately optimize the quantile of outcomes within specific classes of decision rules; Fang et al. (2023) incorporates the quantile of the reward as a regularization term in their objective to maximize the average reward. However, there are no published papers studying quantile policy learning in the offline setting with unobserved confounding.

Causal inference A huge body of works in causal inference has focused on addressing confounding bias through (observed) covariate adjustment (Rubin, 1974; Rosenbaum and Rubin, 1983; Lee et al., 2010; Liu et al., 2024), instrumental variables (Angrist et al., 1996; Ai and Chen, 2003; Newey and Powell, 2003; Chen et al., 2003; Chernozhukov and Hansen, 2008; Chen and Pouzo, 2012; Baiocchi et al., 2014; Hartford et al., 2017), and negative controls (Miao et al., 2018; Kallus et al., 2021). While the covariate adjustment approach cannot address the unobserved confounding, the IV and NC methods can account for the unobserved confounding. While earlier literature on IV and NC focuses on linear objective functions within parametric settings, Chen et al. (2003); Chen and Pouzo (2012, 2015); Chen et al. (2014) and others have studied nonparametric quantile IV, and Miao et al. (2018); Kallus et al. (2021) have considered nonparametric NC with linear objectives. In Theorem C.2 of this paper, we extend the frameworks of Chen and Pouzo (2012) and Kallus et al. (2021) to establish nonlinear functional integral equations of the structural quantile function using NC. In addition, we incorporate the principle of pessimism to address the challenge of limited data coverage in offline quantile policy learning. There are some recent papers that combine pessimism with causal inference tools for confounded offline decision making under linear objective functions (Dong et al., 2023; Chen et al., 2023). Our work extends that of Chen et al. (2023) for maximizing an expected reward under pessimism to optimizing the quantile objective function under pessimism.

In summary, our paper aims at closing an important gap in the existing literature. By deriving integral equations that identify the structural quantile function using IV and NC, and by incorporating a principled pessimism step for offline policy learning with nonlinear objective functions, this work greatly expands the application capabilities of the current literature on offline policy learning.

1.2 Notation

We use upper-case letters to denote random variables and lower-case letters to denote the realizations of the corresponding random variables. For any set $\mathcal{B}$, we let $\Delta(\mathcal{B})$ denote the set of distributions over $\mathcal{B}$. We let p denote the density of random variables and $\mathbb{P}$ denote the probability of an event or probability mass function of a discrete random variable. We use the notation $:=$ to indicate a definition or assignment. We use $\overset{\mathcal{E}}{\leqslant}$ and $\overset{\mathcal{E}}{\geqslant}$ to represent the inequality that holds on some event $\mathcal{E}$. For any n, we let $\mathcal{O}(n)$ denote Cn for some arbitrary positive constant C and $\widetilde{\mathcal{O}}(n)$ to denote $\mathcal{O}(n \cdot \text{poly}(\log n))$.

1.3 Roadmap

In Section 2, we define the offline decision-making problem with a quantile objective when confounding bias is present. We also present several examples motivated from real-world applications. Sections 3 and 4 focus on the IV approach to address unobserved confounding. In Section 3, we first explain the use of IV for the identification and estimation of the causal quantity of interest. We then present two algorithms for IV-assisted quantile-optimal policy learning. In Section 4, we present a thorough theoretical analysis of the regret of our algorithm. Section 5 conducts simulation experiments to illustrate the performance of the algorithm. Section 6 concludes the paper with a short discussion. In §B, we present how our algorithm can be adapted to policy learning under other nonlinear objective functions. §C extends the IV results of Sections 3 and 4 to those using the NC approach.

2 Offline Decision Making

In this section, we introduce the problem of quantile-optimal policy learning within an offline contextual bandit setting, with unmeasured confounders and observed auxiliary variables (IV or NC). We first present the general framework in detail. We then introduce a few real-world motivating examples.

2.1 Offline decision making with a quantile objective

In the following, we formulate the problem of offline decision making with a quantile objective under a contextual bandit framework. Compared with the standard contextual bandits, our setup features the existence of unmeasured confounders and auxiliary variables.

Let $\mathcal{A}$ represent the space of actions and $\Delta(\mathcal{A})$ denote the set of distributions over $\mathcal{A}$. Let $X \in \mathcal{X}$ be the context, $Y \in \mathbb{R}$ the reward, and O the auxiliary variables. Let $U \in \mathcal{U}$ denote the unmeasured confounder that causally affects A and Y simultaneously. The overarching learning problem is to learn a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ that maximizes the average structural quantile of the

reward distribution based on an offline dataset, where the dataset does not include the unmeasured confounders U. Specifically, the decision-making process involves two steps: first, collecting an offline dataset from a given distribution, and then learning a policy from the offline data to apply in an interventional process.

The offline data collection process (ODCP) Suppose there is a joint distribution $p(u, x, o, a, y)$ over the random variables (U, X, O, A, Y), which specifies the joint distribution of the observed data. We note that the joint distribution can be factorized as

$$p(u, x, o, a, y) = p(u, x, o) \cdot p(a \mid u, x, o) \cdot p(y \mid u, x, o, a). \tag{2.1}$$

This factorization illustrates the generating process of the offline data. In particular, $p(u, x, o)$ is the joint distribution of the confounder U, context X, and auxiliary variables O. The term $p(a \mid u, x, o)$ represents the behavior policy, which dictates how actions are selected based on (U, X, O). The reward Y is generated conditioning on (U, X, O, A) via $p(y \mid u, x, o, a)$. Let n denote the number of samples. Let $\{u_i, x_i, o_i, a_i, y_i\}_{i=1}^n$ denote n i.i.d. observations from the joint distribution in (2.1). As U is unmeasured, the offline dataset only includes $\{x_i, a_i, o_i, y_i\}_{i=1}^n$.

We note that, in the presence of unmeasured confounders, the auxiliary variables O should be sufficiently informative for identifying the underlying causal structure. Thus, we need to pose additional assumptions to the joint distribution $p(u, x, o, a, y)$. We will provide two examples of such assumptions in the subsequent subsection, where O is an instrumental variable (Angrist et al., 1996) or a negative control exposure-outcome pair (Lipsitch et al., 2010).

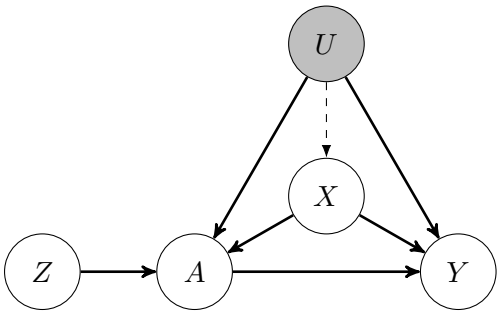
The Interventional process Suppose we learn a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ from the offline dataset, we then apply this policy in an interventional process to assess its performance. In this process, we no longer observe the auxiliary variables. Moreover, the context X may follow a different but known marginal distribution, denoted by $\widetilde{p}(x)$, than that in the ODCP. Upon observing a new context $X \sim \widetilde{p}(x)$ from the environment, we select an action $A \sim \pi(\cdot \mid X)$ from the learned policy, which generates a reward Y from the same confounded data generating process as in the ODCP. Specifically, let p_{in}^{π} denote the joint distribution of random variables (U, X, O, A, Y) in the interventional process under the policy π. We have the following factorization:

$$p_{\text{in}}^{\pi}(u, x, o, a, y) = \widetilde{p}(x) \cdot p(u, o \mid x) \cdot \pi(a \mid x) \cdot p(y \mid u, x, o, a). \tag{2.2}$$

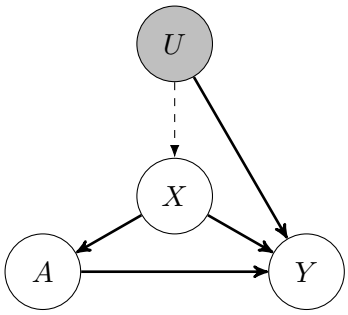
Here $p(y \mid u, x, o, a)$ is the same as in (2.1). Comparing with the joint distribution in the ODCP, we note that the interventional process is a surgical intervention on the ODCP, where we let the action A depend only on the context X. Additionally, the joint distribution of (U, X, O) is allowed to be different. In the job training program example, the interventional process corresponds to the case

where the government uses the learned policy to decide whether to allow a new worker to join the program based on their pre-intervention covariates.

Causal graphs — IV and NC From a causal inference perspective, ODCP and the interventional process can be viewed as two different data generating processes, with two different causal graphs. And the interventional process is a surgical intervention on the ODCP, where we remove the arrows coming into the node of A except the one that comes from the context X. See Figures 1 and 2 for illustrations of the causal graphs of the ODCPs and interventional processes, when O corresponds to an IV Z or a negative control exposure-outcome pair (E, V). Informally, an IV is a random variable that affects the action A but does not directly affect the reward Y. NC are a pair of random variables (E, V), which are called negative control exposure (NCE) and negative control outcome (NCO) respectively. Intuitively, E is an random variable that does not causally affect the reward Y, and V is a random variable causally unaffected by either the action A or the exposure E. We will focus on these causal assumptions in the rest of the paper. To be consistent with the notion in causal inference literature, in the sequel, we replace O with Z when it represents the IV (Angrist et al., 1996), and by (E, V) when it represents the negative control exposure-outcome pair (Lipsitch et al., 2010).



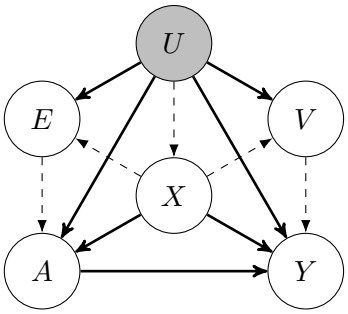
(a) DAG of the ODCP for the IV. Note that $Z \perp\!\!\!\perp Y \mid (X, A, U)$.



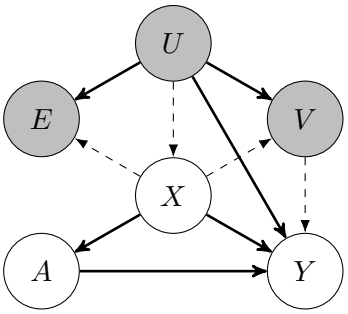
(b) DAG of the interventional process for the IV.

Figure 1: (a) A DAG illustrating the causal relationship between random variables during ODCP when the IV is observed. The dashed edge implies that the causal relationship may be absent. The grey node indicates that U is unmeasured. Here Z is the auxiliary variables. Note that for Z being a valid IV, Z and Y should be independent conditioning on (X, A, U). (b) A DAG encoding the causal relationship between random variables in the interventional process. All arrows coming into the node A have been removed other than the one from the node X.

Quantile objective and regret In the sequel, we denote the expectation over the distribution of random variables in ODCP as $\mathbb{E}$. For any fixed policy π, we denote the expectation over the



(a) A DAG of the ODCP for the NC. Note that we have $V \perp\!\!\!\perp A \,|\, (U, X)$ and $E \perp\!\!\!\perp (Y, V) \,|\, (A, X, U)$.



(b) A DAG of the interventional process for the NC.

Figure 2: (a) A DAG illustrating the causal relationship between random variables of the ODCP when the NCE and the NCO are observed. The dashed edge implies that the causal relationship may be absent. (b) A DAG encoding the causal relationship between random variables in the interventional process. All arrows coming into the node A have been removed other than the one from the node X. Note in addition to U, E and V are also not observed, indicated by the grey nodes.

distribution of random variables in the interventional process as $\mathbb{E}_{p_{\text{in}}^{\pi}}$. We aim to maximize the average structural quantile of the reward distribution. Let $\alpha \in (0, 1)$ be a fixed scalar. Formally, given $X = x$, for any action a, we let $h_{\alpha}^{*}(a, x)$ denote the α-quantile of the potential outcome $Y(a)$, which is called the structural quantile function (Chernozhukov and Hansen, 2008). In the language of do-calculus (Pearl, 1995), we have

$$\mathbb{P}[Y \leqslant h_{\alpha}^{*}(A, X) \,|\, X = x, \text{do}(A = a)] = \alpha \tag{2.3}$$

under the interventional process p_{in}^{π}. We assume that the potential outcome $Y(a)$ given X is continuous; therefore, $h_{\alpha}^{*}(A, X)$ is uniquely defined. Thus, mathematically, our goal is to learn a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ that maximizes the average structural quantile function:

$$v_{\alpha}^{\pi} := \mathbb{E}_{p_{\text{in}}^{\pi}} \left[h_{\alpha}^{*}(A, X) \right], \tag{2.4}$$

where $X \sim \widetilde{p}(x)$ and $A \sim \pi(\cdot \,|\, X)$. Let π^{*} denote the optimal policy that maximizes the average structural quantile function. To measure the performance of a given policy π, we define the regret of π as

$$\text{Regret}(\pi) = v_{\alpha}^{\pi^{*}} - v_{\alpha}^{\pi}. \tag{2.5}$$

The regret is always nonnegative and characterizes the suboptimality of the policy π. Additionally, we would like to highlight that our objective v_{α}^{π} is a nonlinear functional of the distribution of Y, which poses significant challenges to the theoretical analysis.

2.2 Motivating Examples

In many real-world applications, the objective often shifts from maximizing the average outcome to optimizing a specific quantile of the outcome distribution. This shift is particularly relevant when the focus is on distributional fairness or minimizing financial risk. Our framework is designed to address these challenges by enabling quantile optimization, which offers a more equitable approach compared to traditional average-based methods.

Labor market economics Consider the previously mentioned example of job training programs designed to improve the distribution of workers' incomes (Abadie et al., 2002; Chernozhukov and Hansen, 2008). In this context, the auxiliary variable Z_i serves as an IV, which could be whether the worker was invited to participate in the program—this affects their likelihood of participation but does not directly impact their wage. Governments often aim to ensure that the majority of workers experience income gains. Since the average utility can be heavily influenced by a few high-income workers, the focus shifts to optimizing specific quantiles of the utility distribution, such as the median or lower quantiles. This approach helps to ensure that the program benefits a broader range of workers, particularly those at the lower end of the income distribution.

Financial portfolio optimization Consider the design of models for optimizing financial portfolios, where the goal is to enhance portfolio gains. In this setting, let Y_i represent the *gain* (or return) generated by the i-th portfolio over a specific period. The context X_i includes pre-investment characteristics such as market conditions, existing portfolio allocations, and investor risk tolerance. The action A_i denotes the portfolio strategy implemented, for instance, the allocation weights across different asset classes or the choice of specific financial instruments. An IV Z_i might be a factor that influences the chosen strategy A_i but is assumed not to affect the portfolio's gains Y_i directly, except through the strategy itself; examples could include changes in brokerage fee structures or certain types of financial advice that nudge strategy selection without containing direct information about future returns.

Financial institutions often aim for robust performance beyond simply maximizing expected gains. The distribution of portfolio gains can be heavily skewed, particularly by infrequent but large positive or negative events, making the mean gain an unreliable measure of typical performance or risk-adjusted return. Therefore, decision-makers may focus on optimizing specific quantiles of the gain distribution. For instance, they might aim to maximize a lower quantile (e.g., the 5th or 10th percentile gain) to ensure a degree of capital preservation or a minimum performance level, effectively managing downside risk. Alternatively, they might maximize the median gain for a robust measure of central tendency, or a higher quantile (e.g., the 75th percentile) to target ambitious but achievable upside potential.

A key challenge is the potential presence of unmeasured confounders U_i that could affect both the chosen portfolio strategy A_i and the realized gain Y_i. Examples include evolving investor sentiment or private information, which may simultaneously influence asset allocation decisions and market performance, or unobserved market liquidity shocks that drive both strategic adjustments and gain outcomes. Furthermore, adapting concepts like spectral risk measures (Dowd and Blake, 2006) to spectral utility measures, which integrate gains over various quantiles using preference-based weighting functions, is also relevant. In Example B.3, we demonstrate how our method can be adapted to minimize this measure.

Public health policy Consider the scenario of reducing extreme hospital readmission times for patients post-surgery. Here, Y_i represents the post-surgery readmission time for the i-th patient, where lower values are better. A_i represents the type of post-surgery care provided, such as intensive monitoring versus a routine discharge process. X_i is a set of pre-surgery covariates of the patient (e.g., age, comorbidities). To handle unmeasured confounders U_i (e.g., unrecorded severity of the patient's underlying condition that affects both the choice of post-surgery care and readmission risk), NC might be used. For example,

- The Negative Control Exposure (NCE) E_i could be the primary language spoken by the patient's initial care team. It influences communication and thus the nuanced details of care A_i, but is not believed to directly affect the biological process of readmission Y_i, conditional on A_i, X_i, U_i.
- The Negative Control Outcome (NCO) V_i could be a measure of patient satisfaction with hospital administrative processes during admission for the surgery. It is affected by general hospital quality U_i, which also influences readmission time Y_i, but is not affected by the specific post-surgery care A_i or the NCE E_i.

In this context, the goal of the health system is often not just to reduce the average readmission time but to specifically mitigate the risk of very long readmission times (extreme cases), as these often indicate severe adverse events or poor post-surgical outcomes. Thus, the focus is on policy learning that optimizes an upper quantile of the readmission time distribution (e.g., minimizing the 90th or 95th percentile of readmission times).

3 Causal-Assisted Pessimistic Policy Learning Algorithms

The offline quantile policy learning problem introduced in Section 2 involves three challenges: (i) the inherent nonlinearity of the quantile objective as a functional of the reward distribution; (ii) the presence of unmeasured confounders that bias naive estimation; and (iii) insufficient coverage

in the offline dataset. To surmount these obstacles, we develop a suite of algorithms founded on causal inference principles and the pessimism paradigm. Specifically, to address challenges (i) and (ii), we leverage causal inference tools such as instrumental variables (IV) and negative controls (NC) to identify the structural quantile function nonparametrically, which can be solved using a minimax estimation approach to *nonlinear functional integral equations*. This tackles the complication stemming from the nonlinear quantile objective with confounders. Furthermore, to handle challenge (iii), we incorporate the *principle of pessimism*, which leads to the construction of conservative policy estimates. This enables us to focus on policies whose performance can be reliably evaluated with the available offline data.

For clarity, we will first detail these methodologies in the context of IV. The framework for NC, which shares a similar conceptual basis, will be discussed in §C.

3.1 Causal Identification via Instrumental Variables

Optimizing the average structural quantile function $\mathbb{E}_{p_{\text{in}}^{\pi}}[h_\alpha^*(A,X)]$ requires an initial, reliable estimate of the structural quantile function $h_\alpha^*(A,X)$ itself. Recall that $h_\alpha^*(A,X)$ is defined in (2.3). A naive quantile regression of Y on A and X is inadequate due to endogeneity introduced by unmeasured confounders U (Chernozhukov and Hansen, 2008), which leads to a confounding bias.

To address the confounding bias, we propose to use the IV Z to identify $h_\alpha^*(A,X)$. A standard nonparametric IV model requires the IV Z to satisfy three conditions: (i) relevance: the distribution of the action is not constant in the IV, (ii) exclusion restriction: the IV affect the reward only through action, and (iii) unconfounded instrument: the IV are conditionally independent of the error term (Newey and Powell, 2003; Hartford et al., 2017). The general NPQIV model, however, necessitates a more stringent set of assumptions (Chen et al., 2014). Following the treatment of Section 6 in Chen and Pouzo (2012), we employ a simplified version of the NPQIV model incorporating the additive error assumption and conditional moment restriction:

Assumption 3.1 (Model Assumption for Quantile IV)**.** We assume that the following conditions hold for the IV model in the ODCP:

$$Y = h_\alpha^*(A,X) + \epsilon \qquad \text{and} \qquad \mathbb{P}(\epsilon \leqslant 0 \,|\, X, Z) = \alpha. \tag{3.1}$$

Here, ϵ is the structural error, dependent on U, and its α-quantile is zero conditional on X and Z, but not necessarily conditional on X and A.

By Chen et al. (2014), with some regularity conditions, Assumption 3.1 holds by the underlying causal structure of the ODCP in the IV case, as depicted in Figure 1-(a). In contrast to a standard regression model where the mean of the error term is assumed to be zero, we define ϵ as a structural error that adheres to a distinct conditional distribution assumption and is responsible for the

confounding effect induced by the unmeasured confounders U. Consequently, ϵ is not independent of (A, X) and the condition $\mathbb{P}(\epsilon \leqslant 0 \,|\, X, A) = \alpha$ does not hold. This subtlety renders the conventional quantile regression of Y on A and X erroneous. Fortunately, the existence of the IV allows us to write Assumption 3.1 as

$$\mathbb{E}\left[\mathbb{1}\{Y \leqslant h_\alpha^*(A, X)\} \,|\, X, Z\right] = \alpha.$$

If we write $\mathbb{1}\{Y \leqslant h_\alpha^*(A, X)\} - \alpha$ as $W(D; h_\alpha^*)$ where $D = (Y, X, A)$, we would then establish the relationship: $\mathbb{E}\left[W(D; h_\alpha^*) \,|\, X, Z\right] = 0$. That is, h_α^* is the solution to the following conditional moment restriction:

$$\mathbb{E}\left[W(D; h) \,|\, X, Z\right] = 0 \tag{3.2}$$

with respect to h. If h_α^* further is the unique solution to (3.2), we can estimate it by solving the conditional moment restriction in (3.2) based on the offline data.

We note that $W(D, h)$ is a nonlinear function of h and the conditional moment restriction in (3.2) is a nonlinear functional integral equation. For any function h, we define a nonlinear operator $\mathcal{T}^{\mathrm{IV}}$ that maps from a function of (A, X) to a function of (X, Z) as follows:

$$(\mathcal{T}^{\mathrm{IV}} h)(x, z) := \mathbb{E}\left[W(D; h) \,|\, (X, Z) = (x, z)\right].$$

Then we have $\mathcal{T}^{\mathrm{IV}} h_\alpha^* = 0$, i.e., a zero function, almost surely. The equation of the form $\mathcal{T}^{\mathrm{IV}} h = 0$ is referred to as conditional moment restriction in econometrics (Chamberlain, 1992; Ai and Chen, 2003; Newey and Powell, 2003; Chen and Pouzo, 2009). Under the identifiability and realizability assumption (i.e., Assumption 4.1) introduced later, this equation identifies the structural quantile function h_α^*.

As we will show in Theorem C.2, we can also use NC to identify h_α^* under a similar set of assumptions, where we introduce a similar nonlinear operator $\mathcal{T}^{\mathrm{NC}}$. Consequently, estimating h_α^* reduces to solving the conditional moment restrictions $\mathcal{T}^{\mathrm{IV}} h = 0$ or $\mathcal{T}^{\mathrm{NC}} h = 0$. The only distinction between IV and NC is that $\mathcal{T}^{\mathrm{NC}}$ requires solving two conditional moment restrictions simultaneously, while $\mathcal{T}^{\mathrm{IV}}$ involves solving one. This difference is minimal, allowing us to use a similar approach for both cases. In fact, our methodology applies to any case where h_α^* can be identified by solving a conditional moment restriction of the form $\mathcal{T} h = 0$ for some nonlinear operator $\mathcal{T}$. In the sequel, to highlight such generality, we use $\mathcal{T}$ to denote $\mathcal{T}^{\mathrm{IV}}$ for the IV case.

3.2 Minimax Estimation

The conditional moment restriction in (3.2) is essentially a nonparametric inverse problem for h. It is impossible to find a closed-form solution. Various methods have been proposed to estimate the structural quantile function based on the conditional moment restriction. We adopt a minimax

estimation strategy, inspired by Dikkala et al. (2020), to estimate $h^*_\alpha(A, X)$. Our approach transforms the moment restrictions into an unconditional loss function via Fenchel duality and estimates h^*_α by minimizing the loss function.

We first define a conditional residual mean squared error (RMSE) with respect to h as

$$\|\mathcal{T}h\|_2^2 := \mathbb{E}\left[\left(\mathbb{E}\left[W(D;h) \,|\, X, Z\right]\right)^2\right]. \tag{3.3}$$

Note that for any function h we have

$$\|\mathcal{T}h\|_2^2 \geqslant 0, \quad \text{and} \quad \|\mathcal{T}h\|_2^2 = 0 \text{ if and only if } \mathbb{E}\left[W(D;h) \,|\, X, Z\right] = 0.$$

Therefore, we can construct an estimator $h \in \mathcal{H}$ of h^*_α by minimizing (3.3) in a suitable hypothesis function space $\mathcal{H}$. However, (3.3) is a squared conditional expectation. Estimating it using the squared empirical expectation introduces variance terms, leading to a bias. To address this, we first utilize the Fenchel duality of the function $x^2/2$ and reformulate $1/2 \cdot \|\mathcal{T}h\|_2^2$ as

$$\begin{aligned}\frac{1}{2}\|\mathcal{T}h\|_2^2 &= \mathbb{E}\left[\sup_{\theta\in\Theta} \mathcal{T}h(X,Z)\cdot\theta(X,Z) - \frac{1}{2}\cdot|\theta(X,Z)|^2\right] \\ &= \sup_{\theta\in\Theta}\left\{\mathbb{E}\left[W(D;h)\cdot\theta(X,Z) - \frac{1}{2}\cdot|\theta(X,Z)|^2\right]\right\}.\end{aligned} \tag{3.4}$$

Here Θ is a function class over $\mathcal{X}\times\mathcal{Z}$ that contains the function $\mathcal{T}h$, and the last equality holds by the interchangeability principle (Dai et al., 2017). Therefore, we replace the problem of minimizing (3.3) by minimizing (3.4). Note that the expectation in (3.4) is taken with respect to the joint distribution of (Y, A, X, Z) and can be unbiasedly estimated from data. Let $\mathcal{L}_n(h)$ denote the empirical version of (3.4), i.e.,

$$\begin{aligned}\mathcal{L}_n(h) &:= \sup_{\theta\in\Theta}\left\{\mathbb{E}_n\left[W(D;h)\cdot\theta(X,Z)\right] - \frac{1}{2}\|\theta\|_{n,2}^2\right\} \\ &= \sup_{\theta\in\Theta}\left\{\frac{1}{n}\sum_{i=1}^n\left[W(D_i;h)\cdot\theta(X_i,Z_i)\right] - \frac{1}{2n}\sum_{i=1}^n\theta^2(X_i,Z_i)\right\},\end{aligned} \tag{3.5}$$

where we introduce a real-valued test function class Θ on $\mathcal{X}\times\mathcal{Z}$ such that $\mathcal{T}h$ can be well approximated by some function in Θ. As $\mathcal{L}_n(h)$ can now be computed from the offline data, we can then estimate h^*_α by minimizing $\mathcal{L}_n(h)$ with respect to $h \in \mathcal{H}$.

3.3 Two Pessimistic Algorithms for Policy Learning

Based on the loss function $\mathcal{L}_n(h)$ defined in (3.5), we can construct an estimator of h^*_α and use it to learn the optimal policy π^* in the interventional process. We introduce two versions of algorithms based on the pessimism principle.

Challenge of insufficient data coverage. A major challenge of offline decision-making is to deal with the distribution shift between the ODCP that generates the observed action and the oracle policy. To see this, given h and π, we denote $v(h,\pi) := \mathbb{E}_{p_{\text{in}}^{\pi}}[h(A,X)]$. Suppose we have already obtained an estimator $\widehat{h}$ of h_α^* from the offline data, e.g., by minimizing $\mathcal{L}_n(h)$ in (3.5). We can then construct a policy by greedily maximizing $v(\widehat{h},\pi)$, which gives $\widetilde{\pi} := \operatorname{argsup}_\pi v(\widehat{h},\pi)$.

To understand the performance of $\widetilde{\pi}$, we can decompose its regret as

$$\begin{aligned}\text{Regret}(\widetilde{\pi}) &= v_\alpha^{\pi^*} - v_\alpha^{\widetilde{\pi}} \\ &= \underbrace{v_\alpha^{\pi^*} - v(\widehat{h},\pi^*)}_{\text{(i)}} + \underbrace{v(\widehat{h},\pi^*) - v(\widehat{h},\widetilde{\pi})}_{\text{(ii)}} + \underbrace{v(\widehat{h},\widetilde{\pi}) - v_\alpha^{\widetilde{\pi}}}_{\text{(iii)}}. \end{aligned} \tag{3.6}$$

Term (i) only depends on the oracle policy π^* and the estimator $\widehat{h}$. It captures the distribution shift between the interventional process and the ODCP in terms of X. When such a distribution shift is mild and $\widehat{h}$ is a good estimator of h_α^*, this term is small. Term (ii) is upper bounded by zero by the optimality of $\widetilde{\pi}$. Term (iii) presents a unique subtlety: $\widetilde{\pi}$ and $\widehat{h}$ are both derived from offline data, creating spurious correlation. That is, suppose $\widehat{h}$ has a high estimation uncertainty for some suboptimal context-action pair (x,a), causing $\widetilde{\pi}$ to choose this suboptimal action at context x. Term (iii) will be large. In other words, because $\widetilde{\pi}$ is learned from data and is inherently random, to make Term (iii) small, we need to ensure that $\widehat{h}$ is a good estimator of h_α^* for all context-action pairs. Such a requirement is often too strong, especially when some parts of the context-action space are underexplored in the offline data.

The Solution set algorithm. To tackle this challenge, we learn a pessimistic policy by doing uncertainty quantification on $\mathcal{L}_n(h)$ (Yu et al., 2020; Kumar et al., 2020; Kidambi et al., 2020; Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021; Buckman et al., 2020). Specifically, we first construct a solution set $\mathcal{S}(e_n)$ for h based on $\mathcal{L}_n(h)$ as:

$$\mathcal{S}(e_n) := \left\{ h \in \mathcal{H} : \mathcal{L}_n(h) \leqslant \inf_{h\in\mathcal{H}} \mathcal{L}_n(h) + e_n \right\}, \tag{3.7}$$

where e_n is a small positive threshold we will determine later. We will show in Theorem 4.5 (i) that, with high probability, h_α^* lies in $\mathcal{S}(e_n)$ by choosing e_n properly. Then, based on $\mathcal{S}(e_n)$, we select the policy that optimizes the pessimistic average reward function:

$$\widehat{\pi} := \arg\sup_\pi \inf_{h\in\mathcal{S}(e_n)} v(h,\pi). \tag{3.8}$$

Here (3.8) yields a pessimistic solution because $\inf_{h\in\mathcal{S}(e_n)} v(h,\pi) \leqslant v(h_\alpha^*,\pi) = v_\alpha^\pi$ for any π and $h \in \mathcal{S}(e_n)$, when $h_\alpha^* \in \mathcal{S}(e_n)$, which we will later show holds with high probability. In other words, $\widehat{\pi}$ is the policy that maximizes a pessimistic estimate of the objective function.

We state the details of this algorithm in Algorithm 1. To summarize, our algorithm consists of three steps. First, we reduce the problem of estimating h^*_α to the problem of conditional moment restriction in (3.2), which is solved by minimizing the loss function $\mathcal{L}_n(h)$ in (3.4). Second, we construct a solution set (3.7) based on the sublevel sets of the loss function in (3.5). Finally, we construct a policy $\widehat{\pi}$ from the solution set (3.7) by solving the optimization problem in (3.8).

Algorithm 1 Pessimistic Policy Learning Algorithm with Solution Sets

Input: Offline dataset $\{a_i, x_i, z_i, y_i\}_{i=1}^n$, hypothesis space $\mathcal{H}$, test function space Θ and threshold e_n.

(i) Construct solution set $\mathcal{S}(e_n)$ as the sublevel set of $\mathcal{H}$ with respect to metric $\mathcal{L}_n(\cdot)$ and threshold e_n.

(ii) $\widehat{\pi} := \arg\sup_\pi \inf_{h \in \mathcal{S}(e_n)} v(h, \pi)$.

Output: $\widehat{\pi}$.

Benefit of pessimism. To see the benefit of the pessimism principle, recall the regret decomposition in (3.6) for $\widehat{\pi}$ constructed in (3.8). We let $\widehat{h} := \arg\inf_{h \in \mathcal{S}(e_n)} v(h, \widehat{\pi})$. Then, by (3.6), we have

$$v(\widehat{h}, \pi^*) - v(\widehat{h}, \widehat{\pi}) \leqslant 0, \qquad v(\widehat{h}, \widehat{\pi}) - v_\alpha^{\widehat{\pi}} = v(\widehat{h}, \widehat{\pi}) - v(h^*_\alpha, \widehat{\pi}) \leqslant 0.$$

Here, the first inequality is due to the optimality of $\widehat{\pi}$ in (3.8), and the second inequality is due to the fact that $h^*_\alpha \in \mathcal{S}(e_n)$ by our choice of e_n (see Theorem 4.5 (i)). As a result, we have

$$\text{Regret}(\widehat{\pi}) \leqslant v_\alpha^{\pi^*} - v(\widehat{h}, \pi^*) = v\left(h^*_\alpha, \pi^*\right) - v(\widehat{h}, \pi^*). \tag{3.9}$$

Therefore, we reduce the regret of the pessimistic policy $\widehat{\pi}$ to a term that essentially measures the error of the estimator $\widehat{h}$ in the interventional process. In particular, this error is evaluated on the distribution induced by π^*, which is the oracle optimal policy. In other words, as long as the offline data is sufficiently explorative to cover the support of the oracle policy, we can expect the regret of the pessimistic policy $\widehat{\pi}$ to be small.

The Regularized algorithm. As an intermediate step to learn $\widehat{\pi}$, we need to minimize $v(h, \pi)$ for h restricted to $\mathcal{S}(e_n)$. This is an optimization problem with a data-dependent constraint, which is often computationally intractable in practice. We now introduce a computationally benign version of the policy learning algorithm. Let $\mathcal{E}_n(h)$ denote $\mathcal{L}_n(h) - \inf_{h \in \mathcal{H}} \mathcal{L}_n(h)$. We modify the objective function in (3.8) by adding a regularization term $\lambda_n \cdot \mathcal{E}_n(h)$. The regularized version of the pessimistic policy is then defined as:

$$\widehat{\pi}_R := \arg\sup_\pi \inf_{h \in \mathcal{H}} \{v(h, \pi) + \lambda_n \cdot \mathcal{E}_n(h)\}. \tag{3.10}$$

This approach reformulates the data-dependent constrained optimization problem in (3.8) into its augmented Lagrangian counterpart. The idea is motivated from Rashidinejad et al. (2022), which

studies offline bandit and reinforcement learning without confounders. The additional regularization term $\lambda_n \cdot \mathcal{E}_n(h)$ can be viewed as the uncertainty of h. Thus, by adding this term, we effectively penalize the estimator h based on its uncertainty. Furthermore, when the estimator h closely approximates h^*_α, the regularization term approaches $\lambda_n \cdot \mathcal{E}_n(h^*_\alpha)$, thereby becoming a small quantity. Consequently, we expect less bias when optimizing the augmented Lagrangian over π. Thus, the regularization term does not induce significant bias to the optimization problem, while it can significantly penalize the uncertainty of the estimator h. We summarise the regularized version of the algorithm in Algorithm 2.

Comparing (3.10) with (3.8), we see that the regularized version of the pessimistic policy $\widehat{\pi}_R$ is obtained from an unconstrained optimization problem. In practice, this is often more amenable to gradient-based optimization. For example, Xie et al. (2021) uses the mirror descent method to optimize the regularized version of the pessimistic policy. We remark that while Rashidinejad et al. (2022) and Xie et al. (2021) advocate for the utilization of a regularized objective within their respective algorithms, these are explicitly tailored for the domains of offline bandit and offline reinforcement learning without addressing the issue of confounding effect. In contrast, our proposed algorithm is designed to remove the confounding bias, a pervasive challenge of causal inference under the setting of offline decision making.

Algorithm 2 Pessimistic Policy Learning Algorithm with Regularization

Input: Offline dataset $\{a_i, x_i, z_i, y_i\}_{i=1}^n$ from the ODCP, hypothesis space $\mathcal{H}$, test function space Θ and regularization parameter λ_n.

$\widehat{\pi}_R := \arg\sup_\pi \inf_{h\in\mathcal{H}}\{v(h,\pi) + \lambda_n \cdot \mathcal{E}_n(h)\}$.

Output: $\widehat{\pi}_R$.

4 Theoretical Results

In this section, we establish regret guarantees for the pessimistic policy learning algorithms established in (3.8) and (3.10). The theoretical analysis for these algorithms, derived for learning the quantile-optimal policy, is nontrivial and distinct from the existing literature on learning the average-optimal policy for two main reasons. First, the standard concentration inequalities used in the existing literature do not work for $\|\mathcal{T}h\|_2$ because of the nonlinearity of the operator $\mathcal{T}$. Second, it is difficult to bound the regret of $\widehat{\pi}$ or $\widehat{\pi}_R$ in terms of $\|\mathcal{T}h\|_2$.

To address the first challenge, we use bracketing number techniques from van de Geer (2009); Chen et al. (2003). Addressing the second challenge requires us to link the estimation error of the estimated structural quantile function $\widehat{h}$ to $\|\mathcal{T}\widehat{h}\|_2$. That is, we aim to link $\|\widehat{h} - h^*_\alpha\|_\bullet$ to $\|\mathcal{T}\widehat{h}\|_2$, where $\|\cdot\|_\bullet$ is some norm. To see this, consider the estimator $\widehat{\pi}$ defined in (3.8). When the solution set $\mathcal{S}(e_n)$ contains h^*_α and the offline data is sufficiently regular, by the regret decomposition in (3.6),

we can bound Regret($\widehat{\pi}$) by $\|\widehat{h} - h^*_\alpha\|_2$. However, we cannot directly bound $\|\widehat{h} - h^*_\alpha\|_2$ by analyzing the loss function $\mathcal{L}_n(h)$ in (3.5) due to the nonlinearity of the operator $\mathcal{T}$. To bypass this issue, we perform a local expansion of $\mathcal{T}\widehat{h}$ around h^*_α and define a pseudo-metric $\|\cdot\|_{\mathrm{ps}}$ on the solution set $\mathcal{S}(e_n)$. This pseudo-metric acts as a bridge that connects the regret of $\widehat{\pi}$ to $\|\mathcal{T}\widehat{h}\|_2$, which can be bounded using concentration tools. As we show below, although we are focused on the instrumental variable setting, every theorem presented has an equivalent counterpart for negative controls, as detailed in §D. Moreover, the analysis can be extended to general settings where h^*_α is the solution to a nonlinear conditional moment restriction.

In the following, we first present the theoretical guarantees for the solution set algorithm, which yields the pessimistic estimator $\widehat{\pi}$ defined in (3.8). We will focus on the theory of the regularized algorithm in Section 4.4. We introduce several technical assumptions in the sequel and provide a detailed discussion of some of them—specifically, Assumptions 4.1, 4.7, 4.9, 4.10, and Condition 4.4—in §A.

4.1 Analysis of Solution Set $\mathcal{S}(e_n)$

We first establish theoretical guarantees for the solution set $\mathcal{S}(e_n)$ constructed in (3.7). We will prove that, under proper conditions, $\mathcal{S}(e_n)$ contains the true structural quantile function h^*_α with high probability. Moreover, for any function h in $\mathcal{S}(e_n)$, we establish a uniform upper bound on the RMSE $\|\mathcal{T}h\|_2$ defined in (3.3). To begin with, we impose a few conditions on the hypothesis and test function spaces. Note that a function class $\mathscr{F}$ is star-shaped if for every $f \in \mathscr{F}$ and $r \in [0,1]$, we have $rf \in \mathscr{F}$.

Assumption 4.1 (Identifiability and Realizability)**.** We assume $h^*_\alpha \in \mathcal{H}$ and $\mathbb{E}\left[W(D; h^*_\alpha) \,|\, X, Z\right] = 0$. Moreover, for any $h \in \mathcal{H}$ with $\mathbb{E}\left[W(D; h) \,|\, X, Z\right] = 0$, we have $\|h - h^*_\alpha\|_\infty = 0$.

Assumption 4.2 (Compatibility of Test Function Class)**.** We assume that Θ is sufficiently large such that, for any $h \in \mathcal{H}$, $\inf_{\theta\in\Theta} \|\theta - \mathcal{T}h\|_2 = \epsilon_\Theta$ with $\epsilon_\Theta = \mathcal{O}(n^{-1/2})$.

Assumption 4.3 (Regularity of Function Classes)**.** We assume $\mathcal{H}$ is compact with respect to the supremum norm $\|\cdot\|_\infty$ and Θ is star-shaped. We also assume the support of $h^*_\alpha(A, X)$ is bounded, i.e., $\|(h^*_\alpha(A, X))\|_\infty \leqslant L_Y$. Moreover, it holds that $\sup_{h\in\mathcal{H}} \|h\|_\infty \leqslant L_h$ and $\sup_{\theta\in\Theta} \|\theta\|_\infty \leqslant L_\theta$.

Assumption 4.1 states that the conditional moment restriction identifies h^*_α and that the hypothesis space $\mathcal{H}$ captures h^*_α. Moreover, h^*_α is the unique solution to the conditional moment restriction. This set of assumptions is typically required in the econometric literature (Chen and Pouzo, 2012). Assumption 4.2 ensures that the test function class is rich enough to approximate $\mathcal{T}h$ for all $h \in \mathcal{H}$. Assumption 4.3 can be easily satisfied by choosing $\mathcal{H}$ and Θ to be standard uniformly bounded and closed function classes.

By assuming Θ is star-shaped, $\mathcal{L}_n(\cdot)$ becomes a nonnegative function as Θ contains the zero function. Due to the form of $\mathcal{L}_n(\cdot)$, it is natural to consider the function class

$$\mathcal{Q} = \{W(\cdot; h) \cdot \theta(\cdot) : h \in \mathcal{H}, \theta \in \Theta\},$$

which is parameterized by (h, θ). Each element in $\mathcal{Q}$ is a function of (Y, X, A, Z). Since $\mathcal{L}_n(h)$ is represented as the supremum of the addition of two terms, we define an event that quantifies the approximation error between each term in $\mathcal{L}_n(h)$ and its population counterpart. This event will later be proved to have a high probability by using empirical process theory. Fix a sequence of small positive constants η_n that decreases with n. Let $\mathcal{E}$ denote the event

$$\mathcal{E} = \Big\{ \big|\mathbb{E}_n[q(D)] - \mathbb{E}[q(D)]\big| \leqslant \eta_n\Big(\|\theta\|_2 + \eta_n\Big), \text{ and } \big|\|\theta\|_{n,2}^2 - \|\theta\|_2^2\big| \leqslant \tfrac{1}{2}\Big(\|\theta\|_2^2 + \eta_n^2\Big), \ \forall\, q \in \mathcal{Q}\Big\}. \tag{4.1}$$

We impose a condition that ensures $\mathcal{E}$ holds with high probability, which can be verified with concrete instantiations of $\mathcal{H}$ and Θ in §I.

Condition 4.4 . Suppose that Assumption 4.3 holds. For any $\xi > 0$, there exists $\eta_n > 0$ such that the event $\mathcal{E}$ holds with probability at least $1 - 2\xi$.

The set $\mathcal{E}$ encapsulates two events that characterize the concentration of the function classes $\mathcal{Q}$ and Θ separately. The term η_n can be interpreted as the convergence rate of the tail bounds associated with the concentration of $\mathcal{Q}$ and Θ. In cases where $W(\cdot; h)$ is smooth in h, existing literature on the minimax estimation approach (Dikkala et al., 2020; Uehara et al., 2021; Miao et al., 2023) leverages the critical radius of the localized Rademacher complexity of the function classes as developed in Wainwright (2019) to demonstrate that $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$. Nevertheless, given that $W(\cdot; h)$ in our context includes an indicator function parametrized by h, this method is no longer applicable. To achieve a fast convergence rate, we borrow the tools from van de Geer (2009), which measure the complexity of the function class $\mathcal{Q}$ and obtain the concentration bound via the bracketing number. For the detailed computation, we direct the reader to §I, where we establish that Condition 4.4 is satisfied for some $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$ given that $\mathcal{H}$ and Θ are suitably selected. Consequently, Assumption 4.2 can be stated as $\inf_{\theta \in \Theta} \|\theta - \mathcal{T}h\|_2 = \mathcal{O}(\eta_n)$.

The theoretical results for both the solution set and the regularized algorithms are built on the event $\mathcal{E}$. So for the rest of the paper, we assume Condition 4.4 holds. In the following theorem, we prove that, on event $\mathcal{E}$, the solution set $\mathcal{S}(e_n)$ exhibits some favorable properties.

Theorem 4.5 (Uncertainty Quantification)**.** Suppose that Assumptions 4.1, 4.2, and 4.3 hold.

(i). On event $\mathcal{E}$, $\mathcal{L}_n(h_\alpha^*) \leqslant 13/4 \cdot \eta_n^2$. Moreover, if we set $e_n > 13/4 \cdot \eta_n^2$, then $h_\alpha^* \in \mathcal{S}(e_n)$.

(ii). On event $\mathcal{E}$, for all $h \in \mathcal{S}(e_n)$, we have,

$$\|\mathcal{T}h\|_2 = \mathcal{O}\left(\sqrt{e_n}\right) + \mathcal{O}\left(\eta_n\right).$$

Proof. See §E.1 for a detailed proof. □

Theorem 4.5 (i) states that when e_n is chosen properly, the solution set $\mathcal{S}(e_n)$ captures h_α^* with high probability. This fact is indispensable for establishing pessimism, which further leads to the regret upper bound in (3.9). Theorem 4.5 (ii) characterizes the rate of conditional RMSE for all h in $\mathcal{S}(e_n)$ uniformly. As we will show subsequently, this result is crucial for establishing the final regret upper bound of the pessimistic algorithms.

In the following corollary, we apply Theorem 4.5 (i) to the regret decomposition in (3.6) and establish an upper bound of Regret$(\widehat{\pi})$.

Corollary 4.6 (Regret Decomposition)**.** Recall that $\widehat{\pi} := \operatorname{argsup}_\pi \inf_{h \in \mathcal{S}(e_n)} v(h, \pi)$. Suppose that Assumptions 4.1, 4.2 and 4.3 hold. If $e_n > 13/4 \cdot \eta_n^2$, then on the event $\mathcal{E}$, the regret corresponding to $\widehat{\pi}$ is bounded by

$$\text{Regret}(\widehat{\pi}) \leqslant v_\alpha^{\pi^*} - v(\widehat{h}^{\pi^*}, \pi^*) = \mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}^{\pi^*}(A, X)\right]. \tag{4.2}$$

Proof. See §E.4 for a detailed proof. □

4.2 Local Expansion within the Solution Set

The upper bound in Corollary 4.6 involves $\widehat{h}^{\pi^*} - h_\alpha^*$, whose error cannot be directly connected to the RMSE $\|\mathcal{T}h\|_2$ for some h. In the following, we employ a local expansion of $\mathcal{T}\widehat{h}^{\pi^*}$ around h_α^* to bridge $\|\widehat{h}^{\pi^*} - h_\alpha^*\|_\bullet$ for some norm $\|\cdot\|_\bullet$ to $\|\mathcal{T}\widehat{h}^{\pi^*}\|_2$. To this end, we need to first ensure that any element in the solution set $\mathcal{S}(e_n)$ is a consistent estimator of h_α^*.

Before presenting such a result, we first impose a regularity condition on the conditional density of Y given (A, X, Z).

Assumption 4.7 (Regularity of Conditional Density)**.** We assume that the conditional density of Y given (A, X, Z), $p_{Y \mid A,X,Z=(a,x,z)}(y)$ is continuous in (y, a, x, z) and $\sup_y p_{Y \mid A,X,Z}(y) \leqslant K$ for some $K > 0$ for almost all (A, X, Z).

Assumption 4.7 imposes some mild assumptions on the conditional density of Y given (A, X, Z), which is used to ensure that $\mathcal{T}$ is a continuous operator with respect to $\|\cdot\|_\infty$. The assumption is standard in NPQIV (Chen and Pouzo, 2012). We now give a consistent result for any $h \in \mathcal{S}(e_n)$.

Theorem 4.8 (Consistency of functions in $\mathcal{S}(e_n)$)**.** Suppose that Assumptions 4.1, 4.2, 4.3 and 4.7 hold. If $e_n = \mathcal{O}(\eta_n^2)$, then on the event $\mathcal{E}$, for any $h \in \mathcal{S}(e_n)$, we have $\|h - h_\alpha^*\|_\infty = o_p(1)$ and hence $\|h - h_\alpha^*\|_2 = o_p(1)$.

Proof. See §E.2 for a detailed proof. □

Theorem 4.8 shows that if the threshold of the solution set is chosen properly, then any h in $\mathcal{S}(e_n)$ is a consistent estimator of h^*_α. With consistency, one remaining step is required before advancing with local expansion: establishing the correct definition of the derivative of $\mathcal{T}h$ with respect to h. Following Ai and Chen (2012), we introduce the notion of pathwise derivative of $\mathcal{T}h$ evaluated at h^*_α, in the direction $h - h^*_\alpha$, as follows:

$$\frac{d\mathcal{T}h^*_\alpha}{dh}[h - h^*_\alpha] := \frac{d\mathbb{E}[W(D; (1-r)h^*_\alpha + rh) \mid X, Z]}{dr}\Big|_{r=0}.$$

We then define a pseudo-metric on $\mathcal{S}(e_n)$ based on the pathwise derivative:

$$\|h_1 - h_2\|_{\text{ps}} := \sqrt{\mathbb{E}\left[\left(\frac{d\mathcal{T}h^*_\alpha}{dh}[h_1 - h_2]\right)^2\right]}, \tag{4.3}$$

where the term inside the expectation is defined as

$$\frac{d\mathcal{T}h^*_\alpha}{dh}[h_1 - h_2] = \frac{d\mathcal{T}h^*_\alpha}{dh}[h_1 - h^*_\alpha] - \frac{d\mathcal{T}h^*_\alpha}{dh}[h_2 - h^*_\alpha].$$

In other words, for any h_1 and h_2, to compute the pseudo-metric $\|h_1 - h_2\|_{\text{ps}}$, we first compute the pathwise derivative of $\mathcal{T}h^*_\alpha$ evaluated at h^*_α in the direction of $h_1 - h^*_\alpha$ and $h_2 - h^*_\alpha$, and then take the square of the difference of the pathwise derivatives. In particular, suppose $\mathcal{T}$ is a linear operator, e.g., $\mathcal{T}$ can be viewed as a matrix. Then the pseudo-metric $\|h_1 - h_2\|_{\text{ps}}$ reduces to a weighted ℓ_2 norm, i.e., $\|h_1 - h_2\|^2_{\text{ps}} = (h_1 - h_2)^\top(\mathcal{T}^\top\mathcal{T})(h_1 - h_2)$. In Lemma G.1, we derive the closed form of the pseudo-metric for the IV case, which states that

$$\|h - h^*_\alpha\|_{\text{ps}} = \sqrt{\mathbb{E}[(\mathbb{E}[p_{Y|A,X,Z}(h^*_\alpha(A,X))\{h(X,A) - h^*_\alpha(A,X)\} \mid X, Z])^2]}.$$

Let $\mathcal{H}_\epsilon := \{h \in \mathcal{H} : \|h - h^*_\alpha\|_2 \leqslant \epsilon\} \cap \mathcal{S}(e_n)$ be the restricted space of $\mathcal{S}(e_n)$ around the neighborhood of h^*_α where ϵ is a sufficiently small positive number such that $\mathbb{P}(\mathcal{E} \cap \{\widehat{h}^{\pi^*} \in \mathcal{H}_\epsilon\}) \geqslant 1 - 3\xi$. Such an ϵ is guaranteed to exist by Condition 4.4 and Theorem 4.8. We introduce the following assumption on the local curvature of $\mathcal{T}$ around h^*_α.

Assumption 4.9 (Local Curvature)**.** If we set $e_n > 13/4 \cdot \eta_n^2$, then there exists a sufficiently small positive number ϵ and a finite constant $c_0 > 0$ such that for any $h \in \mathcal{H}_\epsilon$, $\|h - h^*_\alpha\|_{\text{ps}} \leqslant c_0 \|\mathcal{T}h\|_2$.

Assumption 4.9 characterizes the local curvature of $\mathcal{T}$. In particular, note that $\mathcal{T}h^*_\alpha = 0$. Thus, under (ii) we have

$$\|h - h^*_\alpha\|_{\text{ps}} \leqslant c_0 \cdot \|\mathcal{T}h - \mathcal{T}h^*_\alpha\|_2,$$

which ensures that the RMSE dominates the pseudo-metric in the neighborhood of h^*_α. The parameter c_0 controls the local curvature of $\mathcal{T}$ around h^*_α. This assumption holds when the conditional moment restriction is locally linear around h^*_α. See the detailed discussion in §A.

4.3 Regret of the Solution Set Offline Policy

Assumption 4.9 allows us to do local expansion and link $\|h - h^*_\alpha\|_{\text{ps}}$ to $\|\mathcal{T}h\|_2$. In the following, we establish the regret of the pessimistic estimator $\widehat{\pi}$ in (3.8) based on the solution set $\mathcal{S}(e_n)$.

Before presenting the theorem, we introduce an assumption on the coverage of the offline data.

Assumption 4.10 (Change of Measure)**.** For the marginal distribution of context $\widetilde{p}$ in the interventional process and the optimal interventional policy π^*, there exists a function $b : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ such that $\mathbb{E}\left[b^2(X, Z)\right] < \infty$ and

$$\mathbb{E}\left[b(X, Z)p_{Y|A,X,Z}(h^*_\alpha(A, X)) \,|\, A = a, X = x\right] = \frac{\widetilde{p}(x)\pi^*(a \,|\, x)}{p(x, a)}. \tag{4.4}$$

Assumption 4.10 can be interpreted as a condition that requires the offline data in ODCP to cover the distribution induced by the oracle policy π^*. The existence of $b(x, z)$ bridges the joint distribution of (X, A) under the ODCP and the one induced by π^*. Recall that the upper bound of regret in Corollary 4.6 involves evaluating the expectation under the interventional policy π^*. The ability to change between measures enables us to evaluate this expectation under the distribution of ODCP. Now we are ready to bound the regret of $\widehat{\pi}$.

Theorem 4.11 (Regret of Solution Set Algorithm)**.** Suppose that structural model Assumption 3.1 holds. Suppose that Assumptions 4.1, 4.2, and 4.3 for function classes $\mathcal{H}$ and Θ hold. Suppose also that Assumptions 4.9 and 4.10 hold. If the threshold e_n for the solution set satisfies $e_n > 13/4 \cdot \eta_n^2$, then, conditioning on the event $\mathcal{E} \cap \{\widehat{h}^{\pi^*} \in \mathcal{H}_\epsilon\}$, the regret of $\widehat{\pi}$ is

$$\text{Regret}(\widehat{\pi}) = c_0 \cdot \|b\|_2 \cdot \mathcal{O}\left(\sqrt{e_n} + \eta_n\right).$$

Proof. See §E.5 for a detailed proof. □

This theorem shows that the regret is bounded by $\mathcal{O}\left(\sqrt{e_n} + \eta_n\right)$ when c_0 and $\|b\|_2$ are regarded as constants. If we set $e_n = \mathcal{O}(\eta_n^2)$, Theorem 4.11 implies that the regret of $\widehat{\pi}$ is bounded by $\mathcal{O}(\eta_n)$. As we will show in Theorem I.7 in §I, $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$ when $\mathcal{H}$ and Θ are chosen to be linear spaces, which corresponds to a "fast statistical rate" (Uehara et al., 2021). This rate aligns with those reported in the existing minimax estimation literature (Dikkala et al., 2020; Li et al., 2022). It is noteworthy that this rate is equivalent to that observed in scenarios involving linear objectives (Chen et al., 2023), where the average reward is maximized rather than the average quantile. This indicates that there is no loss of efficiency when extending the analysis to the nonlinear case.

4.4 Theoretical Analysis for the Regularized Algorithm

For the regularized algorithm, we apply a similar approach using local expansion at h_α^*. Specifically, for any fixed policy π, we define $\widehat{h}_R^\pi$ as

$$\widehat{h}_R^\pi := \arg\inf_{h\in\mathcal{H}}\{v(h,\pi) + \lambda_n \cdot \mathcal{E}_n(h)\}. \tag{4.5}$$

By this definition, we can simplify the regret of $\widehat{\pi}_R$ similar to (4.2) as follows.

Lemma 4.12 (Regret Decomposition)**.** The regret of $\widehat{\pi}_R$ is upper-bounded as follows:

$$\mathrm{Regret}(\widehat{\pi}_R) \leqslant \mathbb{E}_{p_{\mathrm{in}}^{\pi^*}}\left[h_\alpha^*(A,X) - \widehat{h}_R^{\pi^*}(A,X)\right] + \lambda_n \cdot \mathcal{E}_n(h_\alpha^*). \tag{4.6}$$

Proof. See §E.6 for a detailed proof. □

Lemma 4.12 states that the regret of $\widehat{\pi}_R$ can be upper bounded by the average difference between h_α^* and $\widehat{h}_R^{\pi^*}$ in the interventional process, plus a regularization term. Since we define $\mathcal{E}_n(h_\alpha^*)$ as $\mathcal{L}_n(h_\alpha^*) - \inf_{h\in\mathcal{H}}\mathcal{L}_n(h)$, this regularization term is comparatively straightforward to manage. The challenging part, $\mathcal{L}_n(h_\alpha^*)$, has already been addressed in Theorem 4.5 (i). Besides, $\inf_{h\in\mathcal{H}}\mathcal{L}_n(h)$ is the minimum value of the empirical loss function $\mathcal{L}_n(h)$ over the hypothesis set $\mathcal{H}$, which is often sufficiently small when the hypothesis set $\mathcal{H}$ is rich enough. To capture this fact, we impose the following assumption on $\inf_{h\in\mathcal{H}}\mathcal{L}_n(h)$.

Assumption 4.13 (Sample Criterion)**.** We assume that $\inf_{h\in\mathcal{H}}\mathcal{L}_n(h) = \mathcal{O}(\eta_n^2)$.

Assumptions similar to Assumption 4.13 are commonly encountered in the literature of M-estimation (Van der Vaart, 2000). It holds when the empirical loss $\mathcal{L}_n(h)$ can be minimized to a negligible level across the hypothesis set $\mathcal{H}$. This is often satisfied when the optimization method employed in practice is effective. Consequently, by appropriately selecting λ_n, the regularization term mentioned in Lemma 4.12 can be effectively bounded.

It remains to bound the first term on the right-hand side of (4.6). We adopt a similar strategy as in the solution set algorithm. Specifically, we prove that $\widehat{h}_R^\pi$ is consistent to h_α^* with any π. Consequently, we perform a local expansion of $\widehat{h}_R^{\pi^*}$ around h_α^*, allowing the regret to be linked to $\left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2$, which can be further bounded by analyzing the loss function $\mathcal{L}_n$.

The following theorem establishes the consistency of $\widehat{h}_R^\pi$ for h_α^* under the regularized algorithm.

Theorem 4.14 (Consistency of $\widehat{h}_R^\pi$)**.** Suppose that Assumptions 4.1, 4.2, 4.3, 4.7 and 4.13 hold. For any $\epsilon_R > 0$, if we set $\lambda_n \geqslant \eta_n^{-(1+\epsilon_R)}$, then on the event $\mathcal{E}$, for any π, we have $\|\widehat{h}_R^\pi - h_\alpha^*\|_\infty = o_p(1)$ and hence $\|\widehat{h}_R^\pi - h_\alpha^*\|_2 = o_p(1)$.

Proof. See §E.3 for a detailed proof. □

Theorem 4.14 states that on event $\mathcal{E}$, the estimated structural quantile function $\widehat{h}_R^{\pi}$ is consistent to the ground-truth h_α^*. Note the regularization parameter λ_n in Theorem 4.14 only depends on η_n, which characterizes the magnitude of the concentration error. The quantity ϵ_R appears in the definition of λ_n is crucial for the consistency result. Although consistency is theoretically guaranteed for infinitesimally small ϵ_R, the convergence rate of $\widehat{h}_R^{\pi^*}$ to h_α^* becomes slow when ϵ_R is too small. As we will see later, the choice of ϵ_R induces a trade-off in the regret of $\widehat{\pi}_R$. Given the consistency result, we are ready to perform a local expansion of $\widehat{h}_R^{\pi^*}$ around h_α^* in a restricted space of $\mathcal{H}$ around the neighborhood of h_α^*. For a fixed $\lambda_n \geqslant \eta_n^{-(1+\epsilon_R)}$, we fix a sufficiently small positive number ϵ such that the event $\mathcal{E} \cap \left\{\widehat{h}_R^{\pi^*} \in \{h \in \mathcal{H} : \|h - h_\alpha^*\|_2 \leqslant \epsilon\}\right\}$ occurs with probability at least $1 - 3\xi$. Such an ϵ is guaranteed to exist by Condition 4.4 and Theorem 4.14.

Theorem 4.15 (Regret of the Regularized Algorithm)**.** We assume all the assumptions made in Theorem 4.14 hold and additionally impose Assumptions 4.9 and 4.10. Let $0 < \epsilon_R < 1$ be any fixed number. We set the regularization parameter λ_n to $\eta_n^{-(1+\epsilon_R)}$ in (3.10). Then, conditioning on the event $\mathcal{E} \cap \left\{\widehat{h}_R^{\pi^*} \in \{h \in \mathcal{H} : \|h - h_\alpha^*\|_2 \leqslant \epsilon\}\right\}$, the regret of $\widehat{\pi}_R$ is

$$\text{Regret}(\widehat{\pi}_R) = \mathcal{O}\left(\eta_n^{1-\epsilon_R}\right).$$

Proof. See §E.6 for a detailed proof. □

Theorem 4.15 implies that if we set $\lambda_n = \eta_n^{-(1+\epsilon_R)}$, the regret of $\widehat{\pi}_R$ is $\mathcal{O}\left(\eta_n^{1-\epsilon_R}\right)$. Recall by Theorem 4.11, Regret$(\widehat{\pi})$ is $\mathcal{O}(\eta_n)$ when e_n is properly chosen. Thus, the regret of $\widehat{\pi}_R$ is larger than that of $\widehat{\pi}$ by an infinitesimal amount ϵ_R. A small value of ϵ_R slows the convergence of $\widehat{h}_R^{\pi^*}$ toward h_α^*. As a result, the constant c_0 in Assumption 4.9 may be large, which can increase the regret. Essentially, the extra ϵ_R term can be viewed as the trade-off when converting an optimization problem with data-dependent constraints into a nonconstrained optimization problem in the nonlinear quantile case.

Furthermore, when selecting the function spaces $\mathcal{H}$ and Θ as the linear function classes, we have $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$ as shown in Theorem I.7 in §I. This leads to a regret rate of $\widetilde{\mathcal{O}}(n^{-1/2(1-\epsilon_R)})$ for the regularized algorithm, which outperforms the $\widetilde{\mathcal{O}}(n^{-1/3})$ rate associated with the regularized version of the pessimistic offline reinforcement learning algorithm studied in Xie et al. (2021). Since ϵ_R can be infinitesimally small (ignoring c_0), our regret bound also closely matches the $\widetilde{\mathcal{O}}(n^{-1/2})$ regret of the augmented Lagrangian algorithm detailed in Rashidinejad et al. (2022). Note that Rashidinejad et al. (2022) requires the knowledge of the propensity score, which is a strong assumption and is not required by our method. Furthermore, we would like to emphasize that the quantile objective considered in our work is a nonlinear functional of the reward distribution, whereas Xie et al. (2021) and Rashidinejad et al. (2022) focus on the expected reward, which is a linear functional of the reward distribution. To the best of our knowledge, we are the first to establish the efficacy of the

pessimism in the context of nonlinear quantile objectives and unobserved confounding, which is a significant extension of the existing literature on pessimistic offline reinforcement learning.

5 Simulation Experiments

In this section, we evaluate the performance of our approach via numerical simulations. We focus on a synthetic setting where the instrumental variables are available in the offline data. We compare the pessimistic regularized policy learning algorithm, specified in (3.10), against a greedy benchmark that does not incorporate pessimism. The simulation results demonstrate the benefits of pessimism, particularly in scenarios with insufficient data coverage.

5.1 Data Generating Model

We consider a contextual bandit model with two-dimensional contexts, denoted by $X = (X_1, X_2)$, a binary action $A \in \{0, 1\}$, and a one-dimensional instrumental variable Z. The output variable Y is generated according to (3.1), where the structural error ϵ represents an unmeasured confounder that simultaneously influences the action and the outcome. We detail the data generation model below.

Context and instrument generation. We let (X_1, X_2, Z) be drawn from a centered trivariate Gaussian distribution:

$$\begin{bmatrix} X_1 \\ X_2 \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \right),$$

where ρ is a correlation coefficient. We set ρ to 0.95 in all experiments.

Structural error ϵ. For a given target quantile α, we sample the noise ϵ according to $\epsilon \sim \mathcal{N}(-\Phi^{-1}(\alpha), 1)$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal distribution. This construction ensures that the α-quantile of ϵ is zero, i.e., $\mathbb{P}(\epsilon \leqslant 0) = \alpha$.

Action A. Given a treatment assignment score S, which is introduced below, we generate a binary action A from a Bernoulli distribution. In particular, $\mathbb{P}(A = 1 \,|\, S) = \text{sigmoid}(S) = 1/(1 + \exp(-S))$, where $\text{sigmoid}(\cdot)$ is the sigmoid function.

Outcome Y. As shown in (3.10), we generate Y according to the structural model:

$$Y = \beta_1^* \cdot X_1 + \beta_2^* \cdot X_2 + \beta_3^* \cdot A \cdot X_1 + \beta_4^* \cdot A \cdot X_2 + \epsilon, \tag{5.1}$$

where $\beta^* = (\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)^\top = (1, 1, 3, 2)^\top$ are the true coefficients. Here, the term $\beta_1^* \cdot X_1 + \beta_2^* \cdot X_2 + \epsilon$ represents the baseline outcome, and $A \cdot (\beta_3^* \cdot X_1 + \beta_4^* \cdot X_2)$ is the additional effect due to action A.

Treatment assignment score S. The treatment assignment score S is designed to introduce a lack of coverage in the data. Specifically, we introduce two parameters, a structured proportion $p \in [0, 1]$ and two treatment biases b_+ and b_-. For a fraction p of the n samples, we generate S as follows:

$$S = X_1 + 0.5 \cdot X_2^2 + Z + b_+ \cdot \mathbb{1}(X_1 \geqslant 0) + b_- \cdot \mathbb{1}(X_1 < 0) + \epsilon. \tag{5.2}$$

Here we set $b_+ = 8$ and $b_- = -5$. Moreover, for the remaining fraction $(1 - p)$ of the samples, we generate S from random noise: $S \sim \mathcal{N}(0, 1)$.

Our construction induces two layers of complexities in data coverage. First, by having a nonzero p, we introduce out-of-distribution data points that are not generated from the confounded structural model. We set p as 0.7 and 0.8 in our experiments. Second, by having the bias terms b_+ and b_-, we introduce a systematic bias in the treatment assignment. That is, we have more assigned actions for the region $X_1 \geqslant 0$ than for the region $X_1 < 0$.

Optimal policy π^*. By (5.1), it is straightforward to see that the quantile-optimal policy π^* is a deterministic policy that outputs action $A = 1$ if and only if $\beta_3^* \cdot X_1 + \beta_4^* \cdot X_2 > 0$, i.e., $3X_1 + 2X_2 > 0$. When this condition is not satisfied, the optimal action is $A = 0$.

5.2 Algorithm Implementations

We compare the pessimistic regularized policy learning algorithm (Algorithm 2) against a greedy baseline that does not incorporate pessimism.

Linear hypothesis classes $\mathcal{H}$ and Θ. We set $\mathcal{H}$ and Θ to be linear function classes with known features. Specifically, motivated by (5.1), any function h in $\mathcal{H}$ can be written as

$$h(A, X; \beta) = \beta_1 X_1 + \beta_2 X_2 + A(\beta_3 X_1 + \beta_4 X_2), \tag{5.3}$$

where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^\top$ is the parameter vector. The test function space Θ is defined as a linear span of a set of polynomial basis functions involving X_1, X_2, and Z:

$$\{X_1,\ X_2,\ Z,\ X_1X_2,\ X_1Z,\ X_2Z,\ X_1^2,\ X_2^2,\ Z^2,\ X_1^2Z,\ X_1X_2^2,\ Z^3,\ X_1^2X_2^2\}.$$

Greedy approach. This approach first estimates the structural quantile function $h_\alpha(A, X)$ by minimizing the empirical loss function $\mathcal{L}_n(h)$ in (3.5). Note that supremum over θ in (3.5) is a quadratic form, which can be solved exactly in closed form. Thus, we can directly compute the gradient of $\mathcal{L}_n(h)$ with respect to the parameters β of h. We obtain an approximate minimizer $\widehat{h}$ via direct gradient descent, with a constant stepsize 0.05. After obtaining $\widehat{h}$, we derive the greedy policy by solving $\max_\pi v(\widehat{h}, \pi)$, which can be computed in closed form.

Pessimistic approach. We approximately implement Algorithm 2 with $\lambda_n = \sqrt{n}$, where n is the sample size. This choice of λ_n is suggested by Theorem 4.15 for linear function classes, with ϵ_R set to zero. Note that (3.10) involves a minimax optimization and that $\sup_\pi v(h, \pi)$ for any fixed $h \in \mathcal{H}$ can be computed in closed form. We exchange the order of the supremum and infimum in (3.10) and reduce the problem to solving

$$\inf_{h \in \mathcal{H}} \big\{ V(h) + \lambda_n \cdot \mathcal{L}_n(h) \big\}, \qquad \text{where} \quad V(h) = v\big(h, \pi(h)\big). \tag{5.4}$$

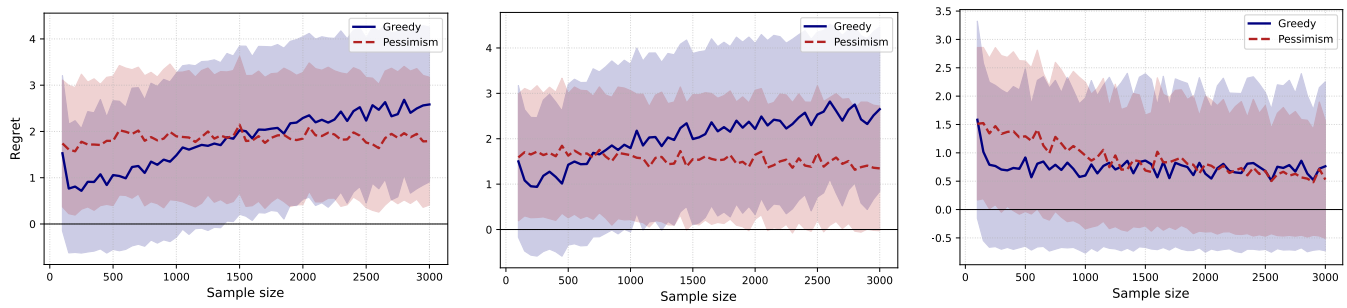
Here we let $\pi(h)$ denote the greedy policy with respect to h. For example, for the function $h(\cdot; \beta)$ defined in (5.3), the greedy policy $\pi(h)$ takes action one if and only if $\beta_3 X_1 + \beta_4 X_2 > 0$.

Another challenge is that $V(h)$ in (5.4) is not differentiable with respect to the parameters of h. Thanks to the low dimensionality of β, to solve this minimization problem, we use a randomized search method. That is, we draw $N = 10^4$ random samples of β from a Gaussian distribution centered at $\widehat{h}$, evaluate the objective values in (5.4) for each sampled β, and select the minimizer $\widetilde{h}$. The greedy policy corresponding to $\widetilde{h}$ is regarded as an approximation of $\widehat{\pi}_R$ in Algorithm 2.

5.3 Experiment Results

We introduce the parameter of structured proportion p in Section 5.1. We compare our pessimistic algorithm against the greedy baseline under different configurations of (n, α, p), where n is the sample size and α is the target quantile. In particular, we choose $n \in \{100, 150, 200, \ldots, 3000\}$, $\alpha \in \{0.15, 0.2, 0.25\}$, and $p \in \{0.7, 0.8\}$. For each configuration, we generate 200 independent datasets and compute the average regret of the learned policies. We show the results for $p = 0.7$ and $p = 0.8$ in Figures 3 and 4, respectively. The blue and red curves represent the average regret of the greedy and pessimistic approaches, respectively, while the shaded areas around the curves indicate one standard deviation above and below the mean. Our findings are summarized as follows.

- In all these figures, the width of the shaded areas is similar for both the greedy and pessimistic approaches, indicating that the variability of the regret is comparable for both methods.
- For the quantile levels $\alpha = 0.15$ and $\alpha = 0.2$, the **pessimistic approach consistently outperforms the greedy approach** when the sample size is large. This clearly demonstrates the advantage of incorporating pessimism, especially when the data coverage is insufficient. In particular, with $(1 - p)$ portion of out-of-distribution data added, it seems to suggest that the signal from the observed data is not sufficient to estimate h^*_α accurately. In this case, pessimism benefits the learning process by avoiding overfitting to the observed data distribution, leading to lower regret.
- Furthermore, it is worth noting that the reason why the greedy approach is inferior is not due to the limited sample size, but rather due to the insufficient coverage of the data generating



(a) $(\alpha, p) = (0.15, 0.7)$ (b) $(\alpha, p) = (0.2, 0.7)$ (c) $(\alpha, p) = (0.25, 0.7)$

Figure 3: Average regret of the greedy and pessimistic approaches for different sample sizes n and quantile levels α, with structured proportion $p = 0.7$. The first two plots clearly shows that pessimism outperforms the greedy approach. The third plot shows that their performances are comparable, with pessimism slightly better when n is large.

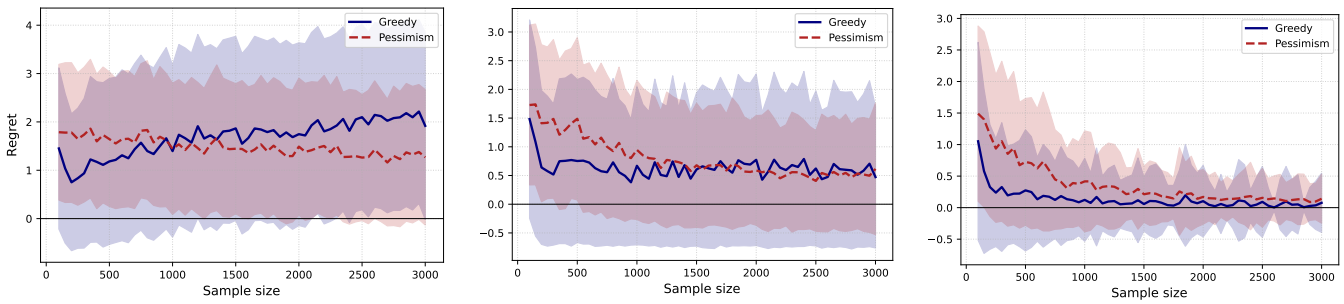
distribution. In particular, as shown in the first two plots of Figures 3 and 4, the margin between the greedy and pessimistic approaches is larger when the sample size is larger.

- For the quantile level $\alpha = 0.25$, pessimism is slightly worse than the greedy approach. In particular, when $p = 0.7$, they are comparable when n is large, as shown in Figure 3-(c). For some sample sizes, pessimism is slightly better. For $p = 0.8$, the greedy approach outperforms the pessimistic approach, as shown in Figure 4-(c). In this case, both algorithms perform well with small regret. This case suggests that when the data coverage is sufficient, greedy algorithm itself can achieve good performance, and the pessimistic approach does not provide significant additional benefits. Moreover, note that the pessimistic approach is computationally more expensive, requiring random sampling in our implementation, and the optimization error can also contribute to the regret. Thus, when the data coverage is sufficient, the greedy approach can be preferred due to its simplicity and efficiency.

- Furthermore, if we compare the performances of these algorithms across different values of p, we observe that both methods perform better when $p = 0.8$ than when $p = 0.7$. This is expected, as a larger structured proportion p implies that the data is more representative of the underlying distribution, leading to better performance for both methods.

6 Discussion

In this work, we propose a quantile-optimal policy learning algorithm where the goal is to find a policy whose reward distribution has the largest α-th quantile for some $\alpha \in (0, 1)$. The idea is



(a) $(\alpha, p) = (0.15, 0.8)$ (b) $(\alpha, p) = (0.2, 0.8)$ (c) $(\alpha, p) = (0.25, 0.8)$

Figure 4: Average regret of the greedy and pessimistic approaches for different sample sizes n and quantile levels α, with structured proportion $p = 0.8$.The first two plots clearly shows that pessimism outperforms the greedy approach. The third plot shows the greedy approach outperforms the pessimistic approach when $\alpha = 0.25$.

to (i) establish a variety of conditional moment restrictions for the identification of the quantile objectives of any policy with the help of instrumental variables and negative controls; (ii) adopt a minimax estimation approach with nonparametric models to solve these conditional moment restrictions, and construct conservative solutions for pessimistic policy learning. The final policy is the one that maximizes the pessimistic objective. In addition, we propose a novel regularized policy learning method that is computationally benign. We deliver a theoretical analysis to show that the proposed method is provably efficient. To the best of our knowledge, this is the first sample-efficient policy learning algorithm for estimating the quantile-optimal policy when there exists unmeasured confounding. Our work requires the observation of either instrumental variables or negative controls. However, we posit that the proposed algorithm is equally effective when the ODCP possesses different underlying causal structures, as long as the quantile treatment effect can be identified by conditional moment restrictions. We leave this investigation to future research.

References

Abadie, A., Angrist, J. and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, **70** 91–117.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 1795–1843.

Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, **170** 442–457.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, **91** 444–455.

Baiocchi, M., Cheng, J. and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine*, **33** 2297–2340.

Buckman, J., Gelada, C. and Bellemare, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*.

Cassel, A., Mannor, S. and Zeevi, A. (2023). A general framework for bandit problems beyond cumulative objectives. *Mathematics of Operations Research*.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society* 567–596.

Chen, S., Wang, Y., Wang, Z. and Yang, Z. (2023). A unified framework of policy learning for contextual bandit with confounding bias and missing observations. *arXiv preprint arXiv:2303.11187*.

Chen, X., Chernozhukov, V., Lee, S. and Newey, W. K. (2014). Local identification of nonparametric and semiparametric models. *Econometrica*, **82** 785–809.

Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71** 1591–1608.

Chen, X. and Pouzo, D. (2009). On nonlinear ill-posed inverse problems with applications to pricing of defaultable bonds and option pricing. *Science in China Series A: Mathematics*, **52** 1157–1168.

Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, **80** 277–321.

Chen, X. and Pouzo, D. (2015). Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, **83** 1013–1079.

Chernozhukov, V. and Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, **73** 245–261.

Chernozhukov, V. and Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, **142** 379–398.

Dai, B., He, N., Pan, Y., Boots, B. and Song, L. (2017). Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*. PMLR.

Dikkala, N., Lewis, G., Mackey, L. and Syrgkanis, V. (2020). Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, **33** 12248–12262.

Dong, J., Mo, W., Qi, Z., Shi, C., Fang, E. X. and Tarokh, V. (2023). Pasta: pessimistic assortment optimization. In *International Conference on Machine Learning*. PMLR.

Dowd, K. and Blake, D. (2006). After var: the theory, estimation, and insurance applications of quantile-based risk measures. *Journal of Risk and Insurance*, **73** 193–229.

Fang, E. X., Wang, Z. and Wang, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, **118** 1733–1746.

Gagliardini, P. and Scaillet, O. (2012). Nonparametric instrumental variable estimation of structural quantile effects. *Econometrica*, **80** 1533–1562.

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*. PMLR.

Hernán, M. A. and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, **183** 758–764.

Horowitz, J. L. and Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, **75** 1191–1208.

Hu, T., Shang, Z. and Cheng, G. (2020). Sharp rate of convergence for deep neural network classifiers under the teacher-student setting. *arXiv preprint arXiv:2001.06892*.

Jin, Y., Yang, Z. and Wang, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*. PMLR.

Kallus, N., Mao, X. and Uehara, M. (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*.

Kidambi, R., Rajeswaran, A., Netrapalli, P. and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, **33** 21810–21823.

Kumar, A., Zhou, A., Tucker, G. and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, **33** 1179–1191.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, **29** 337–346.

Lee, J., Jeon, W., Lee, B., Pineau, J. and Kim, K.-E. (2021). Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*. PMLR.

Lee, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric theory*, **19** 1–31.

Levine, S., Kumar, A., Tucker, G. and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, G., Ma, C. and Srebro, N. (2022). Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *Advances in Neural Information Processing Systems*, **35** 20974–20987.

Li, L., Chu, W., Langford, J., Moon, T. and Wang, X. (2012). An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*. JMLR Workshop and Conference Proceedings.

Linn, K. A., Laber, E. B. and Stefanski, L. A. (2017). Interactive q-learning for quantiles. *Journal of the American Statistical Association*, **112** 638–649.

Lipsitch, M., Tchetgen, E. T. and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, **21** 383.

Liu, M., Ding, L., Yu, D., Liu, W., Kong, L. and Jiang, B. (2022). Conformalized fairness via quantile regression. *Advances in Neural Information Processing Systems*, **35** 11561–11572.

Liu, Q., Chen, Z. and Wong, W. H. (2024). An encoding generative modeling approach to dimension reduction and covariate adjustment in causal inference with observational studies. *Proceedings of the National Academy of Sciences*, **121** e2322376121.

Lu, M., Min, Y., Wang, Z. and Yang, Z. (2022). Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. *arXiv preprint arXiv:2205.13589*.

Metevier, B., Giguere, S., Brockman, S., Kobren, A., Brun, Y., Brunskill, E. and Thomas, P. S. (2019). Offline contextual bandits with high probability fairness guarantees. *Advances in neural information processing systems*, **32**.

Miao, R., Qi, Z., Shi, C. and Lin, L. (2023). Personalized pricing with invalid instrumental variables: Identification, estimation, and policy learning. *arXiv preprint arXiv:2302.12670*.

Miao, W., Geng, Z. and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, **105** 987–993.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, **71** 1565–1578.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82** 669–688.

Prashanth, L., Jagannathan, K. and Kolla, R. K. (2020). Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *Proceedings of the 37th International Conference on Machine Learning*.

Prudencio, R. F., Maximo, M. R. and Colombini, E. L. (2023). A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J. and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, **34** 11702–11716.

Rashidinejad, P., Zhu, H., Yang, K., Russell, S. and Jiao, J. (2022). Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70** 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, **66** 688.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X. and Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W. and Xie, T. (2021). Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*.

van de Geer, S. (2009). *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
https://books.google.com/books?id=0VEcQAAACAAJ

van der vaart, A. and Wellner, J. (2013). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics, Springer New York. https://books.google.com/books?id=zdDkBwAAQBAJ

Van der Vaart, A. W. (2000). *Asymptotic statistics*, vol. 3. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

Wang, L., Yang, Z. and Wang, Z. (2021). Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, **34** 21164–21175.

Wang, Y., Fu, H. and Zeng, D. (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, **113** 1–13.

Wong, W. H. (2022). An equation for the identification of average causal effect in nonlinear models. *Statistica Sinica*, **32** 539–545.

Wu, T. Z., Yu, K. and Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis*, **101** 1607–1621.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P. and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, **34** 6683–6694.

Yang, D., Lafferty, J. and Pollard, D. (2019). Fair quantile regression. *arXiv preprint arXiv:1907.08646*.

Yin, M., Duan, Y., Wang, M. and Wang, Y.-X. (2022). Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C. and Ma, T. (2020). Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, **33** 14129–14142.

Zhu, Q. and Tan, V. (2020). Thompson sampling algorithms for mean-variance bandits. In *International Conference on Machine Learning*. PMLR.

A Discussion of Assumptions and Conditions in Section 4

Assumption 4.1 (Identifiability and Realizability). This assumption, often referred to as a global identification assumption, is commonly made in nearly all nonparametric quantile IV literature. For instance, see Assumption 3.2(ii) in Chen and Pouzo (2012), Assumption 1 in Gagliardini and Scaillet (2012), and Assumption 1 in Horowitz and Lee (2007).

Regarding sufficient conditions for this assumption to hold, Chernozhukov and Hansen (2005) provides such conditions in the case where the context X is discrete. For more general setups, Chen et al. (2014) discusses conditions for local identification in detail. The only study we are aware of that carefully outlines conditions for global identification is Wong (2022). However, Wong's approach relies on a set of strong regularity conditions to establish these results.

Condition 4.4: We only verify that Condition 4.4 holds true for linear function classes because they already encompass a wide variety of functions, including polynomial splines, B-splines, wavelets, and Fourier series basis functions. These function classes can be represented as linear combinations of basis functions, making them instances of the linear function class. By examining our proof in §I, we rely only on the property that linear function classes have a covering number and bracketing number bounded by $A_n \log(C_3/t)$ where A_n is the dimension of the linear function class. Therefore, we can actually obtain different convergence rates η_n under the supremum norm for the chosen function classes. The bracketing and covering numbers of various function classes can be found in different literature. For example, van der vaart and Wellner (2013) provides an extensive treatment of covering numbers and bracketing numbers. They discuss these concepts for a variety of function classes, including: Sobolev spaces, Besov spaces, Hölder classes, and reproducing kernel Hilbert spaces.

Assumption 4.7 (Regularity of Conditional Density). This paper assumes that the outcome variable Y is continuous. The primary reason is that we are dealing with a **nonlinear** functional $\mathcal{T}h$. To link the regret to $\mathcal{T}h$,, we conduct a local expansion of $\mathcal{T}h$ around the neighborhood of the structural quantile function h^*_α. For this expansion, it is essential that the density of Y is continuous and bounded, ensuring that the pathwise derivative $\mathcal{T}h$ at $h = h^*_\alpha$ exists. We remark that this assumption is needed in most of the nonparametric quantile IV literature. For instance, see Condition 6.1(i), (iv) in Chen and Pouzo (2012), Theorem 3.1 in Abadie et al. (2002), Condition A.1 in Chernozhukov and Hansen (2005), and Assumption A.1 in Gagliardini and Scaillet (2012). While our paper does not address discrete outcomes due to this constraint, we believe a similar algorithm could be applied in this case by first convolving the discrete outcome with Gaussian noise, which would smooth the distribution of Y. This smoothing allows for the use of techniques designed for continuous outcomes.

Assumption 4.9 (Local Curvature). Assumption 4.9 asserts that the weaker pseudo metric $\|h - h_\alpha^*\|$ is Lipschitz continuous with respect to the population criterion function $\|\mathcal{T}h\|_2$ within a small neighborhood around h_α^*. This condition is relatively mild and is commonly employed in the nonlinear functional analysis literature (Chen and Pouzo, 2012; Miao et al., 2023). Condition 6.3 (i) in Chen and Pouzo (2012) provides a sufficient condition for the assumption to hold in a broad class of functional spaces $\mathcal{H}$, including the linear function spaces we consider in §I.

Assumption 4.10 (Change of Measure). The Change of Measure Assumption 4.10 is actually a weaker substitute of the standard concentrability assumption commonly used in the RL literature. In the tabular setting, Assumption 4.10 is satisfied when the right-hand side of the assumption is uniformly bounded, and the Moore-Penrose inverse of the probability mass matrix $P(Z \,|\, A, X)^+$ exists. This condition holds if the IV satisfies the standard completeness assumption: $\text{rank}[P(Z \,|\, A, X)] \geqslant |\mathcal{A}| \times |\mathcal{X}|$. We note that Theorem 4.11 and Theorem 4.15 remain valid if Assumption 4.10 is replaced with the usual single-policy concentrability assumption: there exists a constant $\widetilde{c} > 0$ such that $\sup_{x \in \mathcal{X}, a \in \mathcal{A}} \widetilde{p}(x)\pi^*(a \,|\, x)/p(x, a) \leqslant \widetilde{c}$. Note $\widetilde{p}(x)\pi^*(a \,|\, x)/p(x, a)$ represents the density ratio of the distribution under the interventional process with the oracle policy π^* over that of the offline data collection process.

B Other Applications

While the primary focus of this paper is on the general structural quantile model, it's important to note that our algorithm remains applicable as long as the problem can be reformulated into solving a conditional moment restriction. For instrumental variables, the conditional moment restriction $\mathbb{E}\left[W(D; h) \,|\, X, Z\right] = 0$, can be replaced by solving $\mathbb{E}\left[\widetilde{W}(D; h) \,|\, X, Z\right] = 0$, where h is the parameter of interest, related to a functional $\widetilde{W}$ that is specified by the problem. In the case of NC, we can solve (C.4) and (C.5) after substituting $\widetilde{W}(D; h)$ for $W(D; h)$. In particular, our algorithm is effective with more restricted forms of quantile models.

Example B.1 (Single index quantile model (Wu et al., 2010))**.** The structural quantile function is restricted to the form $h(A, X) = h_0(\theta_0^T(X, A))$, where h_0 and θ_0 are unknown. The functional $\widetilde{W}(D; h)$ is then defined as $\mathbb{1}\{Y \leqslant h_0(\theta_0^T(X, A))\} - \alpha$. For this example to fall in our framework, we simply choose $\mathcal{H} = \widetilde{\mathcal{H}} \times \mathbb{R}^{\dim \mathcal{X}} \times \mathbb{R}^{\dim \mathcal{A}}$ to be the function class that contains the function of the form: $h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ defined by $h(X, A) = \widetilde{h}(\alpha_1^T X + \alpha_2^T A)$.

Example B.2 (Partially linear quantile model (Lee, 2003))**.** The structural quantile function is restricted to the form $h(A, X) = h_0(A) + \theta_0^T X$, where h_0 and θ_0 are unknown. The functional $\widetilde{W}(D; h)$ is then defined as $\mathbb{1}\{Y \leqslant h_0(A) + \theta_0^T X\} - \alpha$. We then choose $\mathcal{H} = \widetilde{\mathcal{H}} \times \mathbb{R}^{\dim \mathcal{X}}$, which contains the function of the form: $h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ defined by $h(X, A) = \widetilde{h}(A) + \alpha_1^T X$.

When it is impossible to find a conditional moment restriction, we can still apply our approach in some cases. In Example B.3, we extend the method to a more general setting than maximizing the quantile function, demonstrating the flexibility of the proposed algorithm.

Example B.3 (Quantile-based risk measures (Dowd and Blake, 2006))**.** We consider the spectral risk measure, $\int_0^1 w(\alpha) h^*_\alpha d\alpha$, which is a member of the family of quantile-based risk measures. Here, $w(\alpha)$ is a known weighting function that makes the spectral risk measure coherent. To estimate the spectral risk measure, we consider a finite index set $\mathcal{I} = \{\alpha_1, \ldots, \alpha_m\}$ such that $\alpha_i \in (0,1)$ for any i. We then define

$$M_\phi(A, X; h) := \sum_{i=1}^m \phi(\alpha_i) h_{\alpha_i}(A, X). \tag{B.1}$$

where $h = (h_{\alpha_1}, \ldots, h_{\alpha_m})$, for some known function ϕ. The idea is that we simultaneously estimate $\{h_{\alpha_i}(A, X)\}_{i=1}^m$ for each α_i via the proposed algorithm. Then we approximate $-\int_0^1 w(\alpha) h_\alpha d\alpha$ by $M_\phi(A, X; h)$. The choice of ϕ depends on the numerical method we used to approximate the integral. For example, Newton–Cotes method is one popular choice that has the form of (B.1). Given that $\phi(\alpha_i)$ is fixed, we now introduce the policy-learning algorithm that minimizes the spectral risk measure or, equivalently, maximizes the negative spectral risk measure through the use of IV. The case for NC is similar. For each $i \in [m]$, we define

$$\begin{aligned} W_i(D; h_{\alpha_i}) &= \mathbb{1}\{Y \leqslant h_{\alpha_i}(A, X)\} - \alpha_i, \\ \mathcal{L}_{i,n}(h_{\alpha_i}) &= \sup_{\theta \in \Theta} \left\{ \mathbb{E}_n \left[W_i(D; h_{\alpha_i}) \theta(X, Z) \right] - 1/2 \cdot \|\theta\|^2_{n,2} \right\}. \end{aligned}$$

For the solution set algorithm, we further let

$$\begin{aligned} \mathcal{S}_i(e_n) &= \left\{ h_{\alpha_i} \in \mathcal{H} : \mathcal{L}_{i,n}(h_{\alpha_i}) \leqslant \inf_{h_{\alpha_i} \in \mathcal{H}} \mathcal{L}_{i,n}(h_{\alpha_i}) + e_n \right\}, \text{and} \\ \mathcal{S}(e_n) &= \mathcal{S}_1(e_n) \times \cdots \times \mathcal{S}_m(e_n). \end{aligned}$$

The estimated policy is then given by $\widehat{\pi}^{\mathrm{RM}} := \arg\sup_\pi \inf_{h \in \mathcal{S}(e_n)} \mathbb{E}_{p^\pi_{\mathrm{in}}} [M_\phi(A, X; h)]$. For the regularized algorithm, we define $\mathcal{E}_n(h) := \sum_{i=1}^m \mathcal{L}_{i,n}(h_{\alpha_i}) - \sum_{i=1}^m \inf_{h_{\alpha_i} \in \mathcal{H}} \mathcal{L}_{i,n}(h_{\alpha_i})$. The estimated policy is given by $\widehat{\pi}^{\mathrm{RM}}_R = \mathrm{argsup}_\pi \inf_{h \in \mathcal{H}} \{ \mathbb{E}_{p^\pi_{\mathrm{in}}} [M_\phi(A, X; h)] + \lambda_n \cdot \mathcal{E}_n(h) \}$. We now present a result that characterizes the regret for the algorithms applied in Example B.3.

Corollary B.4. (Regret for Quantile-based Risk Measures). Suppose $\max_{i \in [m]} \phi(\alpha_i) = M$.

(i) Suppose that the assumptions in Theorem 4.11 hold. Specifically, Assumption 4.1 and Assumption 4.3 hold for each $h^*_{\alpha_i}$. Assumption 4.9 holds uniformly for all $h^*_{\alpha_i}$ for some constant c_0. In addition, there exists a change of measure function for each α_i in Assumption 4.10. Then, $\mathrm{Regret}(\widehat{\pi}^{\mathrm{RM}}) = \mathcal{O}(Mm \cdot \eta_n)$ with probability $1 - 3m\xi$.

(ii) Suppose that the assumptions in Theorem 4.15 hold. Specifically Assumption 4.1 Assumption 4.3 hold for each $h^*_{\alpha_i}$. Assumption 4.13 holds for each $\mathcal{L}_{i,n}(h_{\alpha_i})$. Assumption 4.9 holds uniformly for all $h^*_{\alpha_i}$ for some constant c_0. In addition, there exists a change of measure function for each α_i in Assumption 4.10. Then, $\text{Regret}(\widehat{\pi}^{\text{RM}}_R) = \mathcal{O}(Mm \cdot \eta_n^{1-\epsilon_R})$ with probability $1 - 3m\xi$.

Proof. See §E.7 for a detailed proof. □

The results presented in Corollary B.4 align with our intuition: selecting a smaller value for m improves the performance on regret. However, given that m dictates the precision of using (B.1) to estimate the risk measure $\int_0^1 w(\alpha)h^*_\alpha d\alpha$, it essentially represents a trade-off between these two aspects.

C Algorithms and Theory in the Negative Control Case

In this section, we focus on the scenario where the auxiliary variables are negative controls. Unless otherwise specified, the notation in this section is consistent with that used previously. Below, we first introduce the identification assumptions and, for the first time in the literature, establish a novel set of integral equations for the quantile treatment effect using negative controls (Theorem C.2).

C.1 Causal Identification via Negative Controls

During the ODCP, we now observe the negative control exposure (NCE) denoted as E and the negative control outcome (NCO), represented by V. The NCE are variables known to not causally affect the reward Y. The NCO are variables known to be causally unaffected by either the action A or the NCE E. Formally, the following conditions for V and E hold in ODCP:

Assumption C.1 (Negative Controls Assumption)**.** For the action A, unmeasured confounder U, and the context X,

(i) Latent Unconfoundedness: $V \perp\!\!\!\perp A \,|\, (U, X)$ and $E \perp\!\!\!\perp (Y, V) \,|\, (A, X, U)$;

(ii) Completeness Condition: for any $a \in \mathcal{A}, x \in \mathcal{X}$,

$$\mathbb{E}\left[\sigma(U, A, X) \,|\, E = e, A = a, X = x\right] = 0$$

holds for any $e \in \mathcal{E}$ if and only if $\sigma(U, A, X) \overset{\text{a.s.}}{=} 0$. For simplicity, we omit the notation of a.s. in the rest of the paper.

Here, Assumption C.1 (i) formalizes the conditional dependence criteria of the negative controls. Assumption C.1 (ii) is the completeness condition that implies that E captures the variability of

the unmeasured confounders U. This condition appears extensively in the NC literature (Tchetgen et al., 2020). We can now encode the causal structural quantile function into conditional moment restrictions via Assumption C.1.

Theorem C.2 (Conditional Moment Restrictions for Negative Controls)**.** Suppose that Assumption C.1 holds. If there exist bridge functions $h_1^* : \mathcal{A} \times \mathcal{X}$ and $h_2^* : \mathcal{V} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ such that

$$\mathbb{E}\left[W(D; h_1^*) - h_2^*(V, A, X) \,|\, E = e, A = a, X = x\right] = 0, \tag{C.1}$$

$$\mathbb{E}\left[h_2^*(V, a', X) \,|\, X = x\right] = 0, \tag{C.2}$$

for any $(e, a, x, a') \in \mathcal{E} \times \mathcal{A} \times \mathcal{X} \times \mathcal{A}$, then it follows that

$$\mathbb{P}\left[Y \leqslant h_1^*(X, A) \,|\, X = x, \mathrm{do}(A = a)\right] = \alpha.$$

Therefore, the structural quantile function h_1^* is identified by the conditional moment restrictions (C.1) and (C.2). We will henceforth refer to h_1^* and h_α^* interchangeably.

Proof. See §F.1 for a detailed proof. □

Data augmentation In the conditional moment restriction (C.2), the equality holds pointwise for any $a' \in \mathcal{A}$. However, solving this equation becomes computationally intractable if either $|\mathcal{A}|$ is large or the action space $\mathcal{A}$ is continuous. To address this issue, we augment the offline dataset by introducing a random variable A' that is distributed uniformly over $\mathcal{A}$ and is independent of (X, E, V, A, Y). Since

$$\mathbb{E}\left[h_2^*(V, a', X) \,|\, X = x\right] = \mathbb{E}\left[h_2^*(V, A', X) \,|\, A' = a', X = x\right],$$

we can replace (C.2) with

$$\mathbb{E}\left[h_2^*(V, A', X) \,|\, A' = a', X = x\right] = 0. \tag{C.3}$$

In practice, we integrate A' into the offline dataset by appending $a_i' \sim \mathrm{Unif}(\mathcal{A})$ to each sample $(a_i, x_i, e_i, v_i, y_i)$ in ODCP. Consequently, the offline dataset is expanded to $\{a_i, x_i, e_i, v_i, y_i, a_i'\}_{i=1}^n$. We can then estimate h_2^* by solving (C.2) using the augmented offline dataset.

C.2 Estimation via Causal Identification for Negative Controls

Denote $h = (h_1, h_2)$. Based on (C.1) and (C.3), we aim to solve simultaneously for h that satisfies:

$$\mathbb{E}\left[W(D; h_1) - h_2(V, A, X) \,|\, E, A, X\right] = 0 \text{ and} \tag{C.4}$$

$$\mathbb{E}\left[h_2(V, A', X) \,|\, A', X\right] = 0. \tag{C.5}$$

We can define the RMSE on (C.4) and (C.5) with respect to h as

$$\|\mathcal{T}h\|_{2,2}^2 := \mathbb{E}\left[\left(\mathcal{T}_1 h(E,A,X)\right)^2\right] + \mathbb{E}\left[\left(\mathcal{T}_2 h(A',X)\right)^2\right], \tag{C.6}$$

where the operators $\mathcal{T}_1 h$ and $\mathcal{T}_2 h$ denote

$$\mathcal{T}_1 h(\cdot) := \mathbb{E}\left[W(D;h_1) - h_2(V,A,X) \,|\, (E,A,X) = \cdot\right],$$
$$\mathcal{T}_2 h(\cdot) := \mathbb{E}\left[h_2(V,A',X) \,|\, (A',X) = \cdot\right].$$

We then choose the hypothesis function space $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ and the test function space $\Theta = \Theta_1 \times \Theta_2$. Motivated by the Fenchel duality of the function $x^2/2$, we can then minimize the empirical loss function

$$\mathcal{L}_n(h) := \mathcal{L}_{1,n}(h) + \mathcal{L}_{2,n}(h),$$

where we define $\mathcal{L}_{1,n}$ and $\mathcal{L}_{2,n}$ as

$$\mathcal{L}_{1,n}(h) := \sup_{\theta_1 \in \Theta_1} \left\{ \frac{1}{n}\sum_{i=1}^n \{[W(D_i;h_1) - h_2(V_i,A_i,X_i)]\,\theta_1(E_i,A_i,X_i)\} - \frac{1}{2n}\sum_{i=1}^n \theta_1^2(E_i,A_i,X_i) \right\}, \tag{C.7}$$

$$\mathcal{L}_{2,n}(h) := \sup_{\theta_2 \in \Theta_2} \left\{ \frac{1}{n}\sum_{i=1}^n \left[h_2(V_i,A_i',X_i)\right]\theta_2(A_i',X_i) - \frac{1}{2n}\sum_{i=1}^n \theta_2^2(A_i',X_i) \right\}. \tag{C.8}$$

Since both $\mathcal{L}_{1,n}(h)$ and $\mathcal{L}_{2,n}(h)$ can be evaluated from the offline data, we can obtain an estimator of h_1, h_2 by minimizing $\mathcal{L}_n(h) = \mathcal{L}_{1,n}(h) + \mathcal{L}_{2,n}(h)$ over $\mathcal{H}$.

C.3 Algorithm of Policy Learning via Negative Controls

We now present the solution set algorithm for policy learning via NC. We also provide the regularized algorithm in §C.3. The underlying concept of the algorithms closely mirrors that of the IV. The primary caveat is that we are now simultaneously solving two conditional moment equations, with our focus primarily on h_1, as it represents the estimated structural quantile function of interest. Consequently, this necessitates a minor modification to the definition of the solution set. Apart from this adjustment, the foundational ideas remain consistent.

The solution set algorithm for the negative controls We build the solution set for h based on the empirical loss function $\mathcal{L}_n(h)$ and the threshold e_n:

$$\mathcal{S}(e_n) := \left\{ h \in \mathcal{H} : \mathcal{L}_n(h) \leqslant \inf_{h \in \mathcal{H}} \mathcal{L}_n(h) + e_n \right\}. \tag{C.9}$$

As we are only interested in h_1, we additionally define the projection of $\mathcal{S}(e_n)$ onto its first coordinate:

$$\mathcal{S}_1(e_n) := \{h_1 \in \mathcal{H}_1 : \exists (h_1,h_2) \in \mathcal{S}(e_n)\}.$$

We then select the policy π that optimizes the pessimistic average reward function $v(h_1,\pi)$ over the solution set $\mathcal{S}_1(e_n)$. The full algorithm is summarized in Algorithm 3.

Algorithm 3 Quantile action Effect Policy Learning for Negative Controls

Input: Offline dataset $\{a_i, x_i, e_i, v_i, y_i, a_i'\}_{i=1}^n$ from ODCP, hypothesis space $\mathcal{H}$, test function space Θ, and threshold e_n.

(i) Construct the solution set $\mathcal{S}(e_n)$ as the level set of $\mathcal{H}$ with respect to metric $\mathcal{L}_n(\cdot)$ and threshold e_n.

(ii) $\widehat{\pi} := \arg\sup_\pi \inf_{h_1 \in \mathcal{S}_1(e_n)} v(h_1, \pi)$.

Output: $\widehat{\pi}$.

The regularized algorithm for the negative controls Similar to the instrumental variables case, the solution set algorithm faces the computational challenges of solving an optimization problem with data-dependent constraints. We introduce a more practical, regularized version of the algorithm. We denote $\mathcal{E}_n(h) = \mathcal{L}_n(h) - \inf_{h \in \mathcal{H}} \mathcal{L}_n(h)$. The regularized version of the algorithm is summarized in Algorithm 4.

Algorithm 4 The Regularized Policy Learning Algorithm for Negative Controls

Input: Offline dataset $\{a_i, x_i, e_i, v_i, y_i, a_i'\}_{i=1}^n$ from ODCP, hypothesis space $\mathcal{H}$, test function space Θ, and regularization parameter λ_n.

$\widehat{\pi}_R := \arg\sup_\pi \inf_{h_1 \in \mathcal{H}_1} \{v(h_1, \pi) + \lambda_n \mathcal{E}_n(h)\}$.

Output: $\widehat{\pi}_R$.

C.4 Theoretical Results for Policy Learning Algorithm via the Negative Controls

For NC, we are solving two conditional moment equations simultaneously in contrast to solving one equation in the case of IV. Consequently, the underlying assumptions and theoretical results are analogous to those in Section 4. Thus, in this section, we confine our discussion only to the main assumptions and theorems that directly characterize the regret of $\widehat{\pi}$ and $\widehat{\pi}_R$. See §D for a complete derivation of the theoretical results.

A crucial assumption unique to the NC case is Assumption C.3, which parallels the change of measure Assumption 4.10. In this instance, we require the existence of two change of measure functions, with each corresponding to one of the two conditional moment restrictions.

Assumption C.3 (Change of Measure for Negative Controls)**.** For the marginal distribution of context $\widetilde{p}$ in the interventional process and the optimal interventional policy π^*,

(i). There exists a function $b_1 : \mathcal{E} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}\left[b_1^2(E, A, X)\right] < \infty$ and

$$\mathbb{E}\left[b_1(E, A, X) p_{y \mid E,A,X}(h_\alpha^*(A, X)) \mid A = a, X = x\right] = \frac{\widetilde{p}(x)\pi^*(a \mid x)}{p(x, a)}. \tag{C.10}$$

(ii). Define $b_2 : \mathcal{V} \times \mathcal{E} \times \mathcal{A} \times \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ by

$$b_2(v, e, a, x, a') := \frac{\widehat{h}_2^{\pi^*}(v, a, x) p(v \mid e, a, x)}{\widehat{h}_2^{\pi^*}(v, a', x) p(v \mid x)}. \tag{C.11}$$

We assume $\|b_2\|_\infty < \infty$.

Remark on Assumption C.3 Beyond the change of measure condition outlined in Assumption 4.10 for the IV, which requires the function of b to have a finite L_2 norm, Assumption C.3 (ii) incorporates more stringent requirements. First, since we now have to solve two conditional moment restrictions, we now need two change of measure functions, b_1 and b_2. Second, the assumption asserts the change of measure function b_2 to have a finite L_∞ norm. This is a stricter condition as L_∞ is a stronger norm than L_2. Practically, this implies the necessity for the offline dataset to uniformly cover the covariate space $\mathcal{X}$ and the negative control outcome space V. In addition, to ensure that $\widehat{h}_2^{\pi^*}(v, a', x)$ is bounded away from 0, we can do a location shift on the reward.

Theorem C.4 (Informal Version of Theorem D.9)**.** Under appropriate conditions, if the threshold e_n for the solution set is set to $e_n > (2L_h^2 + 5/4)\eta_n^2$, then with probability $1 - 5\xi$, the regret of $\widehat{\pi}$ is

$$\text{Regret}(\widehat{\pi}) = c_1(1 + \|b_2\|_\infty)\|b_1\|_2 \cdot \left(\mathcal{O}\left(\sqrt{e_n}\right) + \mathcal{O}\left(\eta_n\right)\right).$$

Proof. See Theorem D.9 for the complete statement of the theorem and §F.2 for a detailed proof. □

Theorem C.5 (Informal Version of Theorem D.12)**.** Under appropriate conditions, for any $0 < \epsilon_R < 1$, if the regularized parameter λ_n is set to $\lambda_n = \eta_n^{-(1+\epsilon_R)}$, then with probability $1 - 5\xi$, the regret of $\widehat{\pi}_R$ is

$$\text{Regret}(\widehat{\pi}_R) = \mathcal{O}\left(\eta_n^{1-\epsilon_R}\right).$$

Proof. See Theorem D.12 for the complete statement of the theorem and §F.4 for a detailed proof. □

D Theoretical Analysis for Algorithm via Negative Controls

The structure of this section is parallel to that of Section 4. The ultimate goal is to link the regret of $\widehat{\pi}$ and $\widehat{\pi}_R$ to $\|\mathcal{T}h\|_{2,2}^2$. We first construct two events of high probability that bridge the empirical loss $\mathcal{L}_n$ to $\|\mathcal{T}h\|_{2,2}^2$.

$$\begin{aligned}
\widetilde{\mathcal{E}}_1 := \Big\{ &\left|\mathbb{E}_n\left\{[W(D;h_1) - h_2(V,A,X)]\,\theta_1(E,A,X)\right\} - \mathbb{E}\left\{[W(D;h_1) - h_2(V,A,X)]\,\theta_1(E,A,X)\right\}\right| \\
&\leqslant \eta_n\left(\|\theta_1\|_2 + \eta_n\right), \quad \left|\|\theta_1\|_{n,2}^2 - \|\theta_1\|_2^2\right| \leqslant \frac{1}{2}\left(\|\theta_1\|_2^2 + \eta_n^2\right), \forall h \in \mathcal{H}, \forall \theta_1 \in \Theta_1\Big\}. \\
\widetilde{\mathcal{E}}_2 := \Big\{ &\left|\mathbb{E}_n\left\{\left[h_2(V,A',X)\right]\theta_2(V,A',X)\right\} - \mathbb{E}\left\{\left[h_2(V,A',X)\right]\theta_2(V,A',X)\right\}\right| \\
&\leqslant \eta_n\left(\|\theta_2\|_2 + \eta_n\right), \quad \left|\|\theta_2\|_{n,2}^2 - \|\theta_2\|_2^2\right| \leqslant \frac{1}{2}\left|\|\theta_2\|_2^2 + \eta_n^2\right|, \forall h \in \mathcal{H}, \forall \theta_2 \in \Theta_2\Big\}.
\end{aligned}$$

Assumption D.1 (Regularity of Function Classes for Negative Controls)**.** We assume $\mathcal{H}$ is compact with respect to the norm $\|\cdot\|_{\sup}$ and Θ is star-shaped. We also assume $\sup_{h\in\mathcal{H}} \|h\|_{\sup} \leqslant L_H$.

Condition D.2 . Suppose that Assumption D.1 holds. For any $\xi > 0$, there exists $\eta_n > 0$ that decreases with n such that the event $\widetilde{\mathcal{E}} = \widetilde{\mathcal{E}}_1 \cap \widetilde{\mathcal{E}}_2$ holds with probability at least $1 - 4\xi$.

Note each $\widetilde{\mathcal{E}}_i$ bridges $\mathcal{T}_i h(\cdot)$ to $\|\mathcal{T}_i h\|_2^2$. Hence we denote $\widetilde{\mathcal{E}} = \widetilde{\mathcal{E}}_1 \cap \widetilde{\mathcal{E}}_2$, so that $\widetilde{\mathcal{E}}$ is an event that controls the difference between the empirical loss function $\mathcal{L}_n(h)$ and $\|\mathcal{T}h\|_{2,2}^2$. If we choose $\mathcal{H}_1, \mathcal{H}_2, \Theta_1$ and Θ_2 to be linear function classes, Condition D.2 holds for some $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$. The reason is that by applying a similar strategy as in §I, we can show that each $\widetilde{\mathcal{E}}_i$ holds with probability at least $1 - 2\xi$. Taking the union bound gives us $1 - 4\xi$. We define a norm $\|\cdot\|_{\sup}$ on $\mathcal{H} := \mathcal{H}_1 \times \mathcal{H}_2$ as

$$\|h\|_{\sup} := \|h_1\|_\infty + \|h_2\|_\infty .$$

Assumption D.3 (Identifiability and Realizability for the Negative Controls)**.** Suppose $(h_1^*, h_2^*) \in \mathcal{H}$. For any $h \in \mathcal{H}$ that satisfies (C.4) and (C.5), we have $\|h - h^*\|_{\sup} = 0$.

Assumption D.3 ensures that the structural quantile function is uniquely identified through the conditional moment restrictions. Consequently, we can substitute h_1^* with h_α^* without any ambiguity.

Assumption D.4 (Compatibility of Test Function Class)**.** For any $h \in \mathcal{H}$, $\inf_{\theta\in\Theta} \|\theta - \mathcal{T}h\|_2 = \epsilon_\Theta$, and $\epsilon_\Theta = \mathcal{O}(n^{-1/2})$.

D.1 Solution Set Algorithm for the Negative Controls

We introduce an analogue to Theorem 4.5 tailored for NC. This demonstrates that within the event $\widetilde{\mathcal{E}}$, the solution set $\mathcal{S}_1(e_n)$ exhibits several favorable properties.

Theorem D.5 (Uncertainty Quantification for Negative Controls)**.** Suppose that Assumptions D.3, D.4, and D.1 hold.

(i). On event $\widetilde{\mathcal{E}}$, $\mathcal{L}_n(h^*) \leqslant (2L_h^2 + 5/4)\eta_n^2$ where $h^* = (h_\alpha^*, h_2^*)$. Moreover, if we set $e_n > (2L_h^2 + 5/4)\eta_n^2$, then $h_\alpha^* \in \mathcal{S}_1(e_n)$.

(ii). On event $\widetilde{\mathcal{E}}$, for all $h \in \mathcal{S}_1(e_n)$, we have,

$$\|\mathcal{T}h\|_2 = \mathcal{O}\left(\sqrt{e_n}\right) + \mathcal{O}\left(\eta_n\right).$$

Proof. The proof is identical to the proof of Theorem 4.5, which is in §E.1. □

Assumption D.6 (Regularity of Density for the Negative Controls)**.** We assume that $p_{y\,|\,e,a,x}$, the conditional density of Y given (E, A, X), exists. Moreover, $p_{y\,|\,e,a,x}(y)$ is continuous in (y, e, a, x) and $\sup_y p_{y\,|\,e,a,x}(y) < \infty$ for almost all (y, e, a, x).

Theorem D.7 (Consistency of Functions in $\mathcal{S}(e_n)$ for Negative Controls)**.** Suppose that Assumptions D.3, D.4, D.1, and D.6 hold. If we set $e_n = \mathcal{O}(\eta_n^2)$, then on the event $\widetilde{\mathcal{E}}$, for any $h \in \mathcal{S}(e_n)$, $\|h_1 - h_\alpha^*\|_\infty = o_p(1)$ and hence $\|h_1 - h_\alpha^*\|_2 = o_p(1)$.

Proof. Given Lemma H.3, the proof is identical to that of Theorem 4.8, which is in §E.2. □

Assumption D.6 ensures that the conditional density of the reward Y given (E, A, X) is well-defined and bounded away from infinity. Note we only present the result of consistency of h_1 in theorem 4.8. This focus is deliberate, as the upper bound of regret of the NC case is the same as that detailed in Corollary 4.6. Thus, the upper bound only concerns the conditional average difference between h_1 and h_α^*. Therefore, a local expansion of h_1 around h_α^* suffices to serve our purpose.

let $\mathcal{H}_{1\epsilon} := \{h_1 \in \mathcal{H}_1 : \|h_1 - h_\alpha^*\|_2 \leqslant \epsilon\} \cap \mathcal{S}_1(e_n)$ where ϵ is a sufficiently small positive number such that $\mathbb{P}\left(\widetilde{\mathcal{E}} \cap \left\{\widehat{h}_1^{\pi *} \in \mathcal{H}_{1\epsilon}\right\}\right) \geqslant 1 - 5\xi$. Such ϵ is guaranteed to exist by Condition D.2 and Theorem D.7.

Assumption D.8 (Local Curvature for the Negative Controls)**.** If we set $e_n > (2L_h^2 + 5/4)\eta_n^2$, then there exists a finite constant $c_1 > 0$ such that for any $h_1 \in \mathcal{H}_{1\epsilon}$,

$$\|h_1 - h_\alpha^*\|_{\text{ps}} \leqslant c_1 \|\mathbb{E}[W(D; h_1) \,|\, E, A, X]\|_2 .$$

Similar to Lemma G.1, we can show that

$$\|h_1 - h_\alpha^*\|_{\text{ps}} = \sqrt{\mathbb{E}[(p_{y\,|\,E,A,X}(h_\alpha^*(A, X))\,\{h_1(X, A) - h_\alpha^*(A, X)\})^2]}.$$

Theorem D.9 (Regret of Solution Set Algorithm for Negative Controls)**.** Suppose that Assumption C.1 holds. Suppose that Assumptions D.3, D.4, and D.1 for function classes $\mathcal{H}$ and Θ hold. Suppose also that Assumptions D.8 and C.3 hold. If the threshold e_n for the solution set is set to $e_n > (2L_h^2 + 5/4)\eta_n^2$, then conditioning on the event $\widetilde{\mathcal{E}} \cap \left\{\widehat{h}_1^{\pi *} \in \mathcal{H}_{1\epsilon}\right\}$, the regret of $\widehat{\pi}$ is

$$\text{Regret}(\widehat{\pi}) = c_1(1 + \|b_2\|_\infty)\,\|b_1\|_2 \cdot \left(\mathcal{O}\left(\sqrt{e_n}\right) + \mathcal{O}\left(\eta_n\right)\right).$$

Proof. See §F.2 for a detailed proof. □

D.2 Regularized Algorithm for Negative Controls

We now detail the theoretical analysis for the regularized algorithm for NC. We define $\widehat{h}_R^\pi := (\widehat{h}_{R1}^\pi, \widehat{h}_{R2}^\pi) := \arg\inf_{h \in \mathcal{H}} \{v(h_1, \pi) + \lambda_n \cdot \mathcal{E}_n(h)\}$.

Assumption D.10 (Sample Criterion for Negative Controls)**.** We assume $\inf_{h \in \mathcal{H}} \mathcal{L}_n(h) = \mathcal{O}(\eta_n^2)$.

Theorem D.11 (Consistency of $\widehat{h}_R^{\pi}$ for the Negative Controls)**.** Suppose that Assumptions D.3, D.4, D.1, D.6 and D.10 hold. For any $\epsilon_R > 0$, if we set $\lambda_n \geqslant \eta_n^{-(1+\epsilon_R)}$, then on the event $\widetilde{\mathcal{E}}$, for any π, $\|\widehat{h}_{R1}^{\pi} - h_\alpha^*\|_\infty = o_p(1)$ and hence $\|\widehat{h}_{R1}^{\pi} - h_\alpha^*\|_2 = o_p(1)$.

Proof. See §F.3 for a detailed proof. □

Given the consistency result, we then perform local expansion of $\widehat{h}_{R1}^{\pi^*}$ around h_α^*. We now define a restricted space of $\mathcal{H}$. For a fixed $\lambda_n \geqslant \eta_n^{-(1+\epsilon_R)}$, let $\mathcal{H}_{1\epsilon} := \{h_1 \in \mathcal{H}_1 : \|h_1 - h_\alpha^*\|_2 \leqslant \epsilon\}$. We then fix a sufficiently small positive number ϵ such that the event $\widetilde{\mathcal{E}}_{R\epsilon} = \widetilde{\mathcal{E}} \cap \{\widehat{h}_{R1}^{\pi^*} \in \mathcal{H}_{1\epsilon}\}$ occurs with probability at least $1 - 5\xi$ by Condition D.2.

Theorem D.12 (Regret of the Regularized Algorithm for the Negative Controls)**.** Suppose that the conditions for negative controls Assumption C.1 hold. Suppose that Assumptions D.3, D.4, and D.1 for function classes $\mathcal{H}$ and Θ hold. Suppose also that the Assumptions D.10, D.8, and C.3 hold. For any $0 < \epsilon_R < 1$, if the regularized parameter λ_n is set to $\lambda_n = \eta_n^{-(1+\epsilon_R)}$, then conditioning on the event $\widetilde{\mathcal{E}}_{R\epsilon}$, the regret of $\widehat{\pi}_R$ is

$$\text{Regret}(\widehat{\pi}_R) = \mathcal{O}\left(\eta_n^{1-\epsilon_R}\right).$$

Proof. See §F.4 for a detailed proof. □

E Proof of the Main Results of Section 4

In this section, we prove the main results in Section 4. The goal is to establish the regret of the solution set algorithm and the regularized algorithm for the IV case.

E.1 Proof of Theorem 4.5

Theorem 4.5 closely parallels Theorem 6.4 in Chen et al. (2023), with adaptations made to suit our specific context. Despite the resemblance of the theorem statement and proof techniques, we nevertheless include a detailed proof to accommodate the shift from a linear operator to a nonlinear operator, which results in several changes in how the final results are presented. In particular, with Assumptions 4.1 and 4.2, we prove that $h_\alpha^* \in \mathcal{S}(e_n)$ by deriving an upper bound for $\mathcal{L}_n(h_\alpha^*)$. We then derive the upper bound for $\|\mathcal{T}h\|_2$ for any $h \in \mathcal{S}(e_n)$.

Theorem 4.5 (i). In this part, we first compute an upper bound of $\mathcal{L}_n(h_\alpha^*)$. We then deduce that $h_\alpha^* \in \mathcal{S}(e_n)$ if $e_n > 13/4 \cdot \eta_n^2$.

Proof. It holds on $\mathcal{E}$ that

$$\begin{aligned}
\mathcal{L}_n(h_\alpha^*) &= \sup_{\theta\in\Theta} \mathbb{E}_n\left[W(D;h_\alpha^*)\theta(Z,X)\right] - \frac{1}{2}\left\|\theta\right\|_{n,2}^2 \\
&\leqslant \sup_{\theta\in\Theta}\Big\{ \left|\mathbb{E}_n\left[W(D;h_\alpha^*)\theta(Z,X)\right] - \mathbb{E}\left[W(D;h_\alpha^*)\theta(Z,X)\right]\right| \\
&\quad + \frac{1}{2}\left|\left\|\theta\right\|_{n,2}^2 - \left\|\theta\right\|_2^2\right| + \mathbb{E}\left[W(D;h_\alpha^*)\theta(Z,X)\right] - \frac{1}{2}\left\|\theta\right\|_2^2 \Big\} \\
&\overset{\mathcal{E}}{\leqslant} \sup_{\theta\in\Theta}\Big\{\eta_n\left(\left\|\theta\right\|_{n,2} + \eta_n\right) + \frac{1}{4}\left(\left\|\theta\right\|_2^2 + \eta_n^2\right) \\
&\quad + \mathbb{E}\left[W(D;h_\alpha^*)\theta(Z,X)\right] - \frac{1}{2}\left\|\theta\right\|_2^2\Big\},
\end{aligned}$$

where the first inequality holds by the triangle inequality and the second inequality holds by the definition of $\mathcal{E}$. Let $\mathcal{L}^\lambda(\cdot) = \sup_{\theta\in\Theta}\mathbb{E}\left[W(D;h_\alpha^*)\theta(Z,X)\right] - \lambda\left\|\theta\right\|_2^2$. Then $\mathcal{L}_n(h_\alpha^*)$ satisfies

$$\mathcal{L}_n(h_\alpha^*) \leqslant \mathcal{L}^{1/8}(h_\alpha^*) - \inf_{\theta\in\Theta}\left(\frac{1}{8}\left\|\theta\right\|_2^2 - \eta_n\left\|\theta\right\|_2\right) + \frac{5}{4}\eta_n^2. \tag{E.1}$$

We further let $\theta^\lambda(\cdot;h) := \arg\sup_{\theta\in\Theta}\mathcal{L}^\lambda(h)$. To relate $\mathcal{L}^{1/8}$ to $\mathcal{L}^{1/2}$, we observe that for $\mathcal{L}^{\lambda_1}(h)$ and $\mathcal{L}^{\lambda_2}(h)$ where $0 < \lambda_1 \leqslant \lambda_2$,

$$\begin{aligned}
\mathcal{L}^{\lambda_2}(h) &= \sup_{\theta\in\Theta}\mathbb{E}\left[W(D;h)\theta(Z,X)\right] - \lambda_2\left\|\theta\right\|_2^2 \\
&= \frac{\lambda_2}{\lambda_1}\cdot\sup_{\theta\in\Theta}\left\{\mathbb{E}\left[\frac{\lambda_1}{\lambda_2}W(D;h)\theta(Z,X)\right] - \lambda_1\left\|\theta\right\|_2^2\right\} \\
&\geqslant \frac{\lambda_2}{\lambda_1}\cdot\left(\mathbb{E}\left[\frac{\lambda_1}{\lambda_2}W(D;h)\cdot\frac{\lambda_1}{\lambda_2}\theta^{\lambda_1}(Z,X;h)\right] - \lambda_1\left\|\frac{\lambda_1}{\lambda_2}\theta^{\lambda_1}(Z,X;h)\right\|_2^2\right) \\
&\geqslant \frac{\lambda_1}{\lambda_2}\mathcal{L}^{\lambda_1}(h),
\end{aligned} \tag{E.2}$$

where the first inequality holds by letting $\theta = \lambda_1/\lambda_2\cdot\theta^{\lambda_1}$. Note $\frac{\lambda_1}{\lambda_2}\theta^{\lambda_1}\in\Theta$ as $\lambda_1\leqslant\lambda_2$ and Θ is star-shaped. Now substituting $\lambda_1 = 1/8$ and $\lambda_2 = 1/2$ and plugging (E.2) into (E.1):

$$\begin{aligned}
\mathcal{L}_n(h_\alpha^*) &\overset{\mathcal{E}}{\leqslant} 4\mathcal{L}^{1/2}(h_\alpha^*) - \inf_{\theta\in\Theta}\left(\frac{1}{8}\left\|\theta\right\|_2^2 - \eta_n\left\|\theta\right\|_2\right) + \frac{5}{4}\eta_n^2 \\
&\leqslant 4\mathcal{L}^{1/2}(h_\alpha^*) + \frac{13}{4}\eta_n^2,
\end{aligned} \tag{E.3}$$

where the second inequality holds by computing the minimum value of the quadratic equation of $\left\|\theta\right\|_2^2$. Also note that

$$\begin{aligned}
\mathcal{L}^{1/2}(h_\alpha^*) &= \sup_{\theta\in\Theta}\left\{\mathbb{E}\left[W(D;h_\alpha^*)\theta(Z)\right] - \frac{1}{2}\left\|\theta\right\|_2^2\right\} \\
&\leqslant \sup_{\theta}\left\{\mathbb{E}\left[W(D;h_\alpha^*)\theta(Z)\right] - \frac{1}{2}\left\|\theta\right\|_2^2\right\} \\
&= \frac{1}{2}\left\|\mathcal{T}h_\alpha^*\right\|_2^2,
\end{aligned}$$

where the second inequality follows by Fenchel duality. Hence (E.3) gives

$$\mathcal{L}_n(h_\alpha^*) \overset{\mathcal{E}}{\leqslant} 2\left\|\mathcal{T}h_\alpha^*\right\|_2^2 + \frac{13}{4}\eta_n^2 = \frac{13}{4}\eta_n^2. \tag{E.4}$$

By the non-negativity of $\mathcal{L}_n(\cdot)$ [1], it follows that

$$\mathcal{L}_n(h_\alpha^*) - \inf_{h\in\mathcal{H}} \mathcal{L}_n(h) \overset{\mathcal{E}}{\leqslant} \frac{13}{4}\eta_n^2.$$

Therefore, by definition of the solution set in (3.7), with $e_n > 13/4 \cdot \eta_n^2$, it holds on $\mathcal{E}$ that $h_\alpha^* \in \mathcal{S}(e_n)$. □

Theorem 4.5 (ii). In this part, we derive an upper bound for $\|\mathcal{T}h\|_2^2$ for any $h \in \mathcal{S}(e_n)$.

Proof. We first note it holds for all $h \in \mathcal{S}(e_n)$ that

$$\begin{aligned}
\mathcal{L}_n(h) &= \sup_{\theta\in\Theta} \mathbb{E}_n\left[W(D;h)\theta(Z,X)\right] - \frac{1}{2}\|\theta\|_{n,2}^2 \\
&\geqslant \sup_{\theta\in\Theta}\Big\{ -\left|\mathbb{E}_n\left[W(D;h)\theta(Z,X)\right] - \mathbb{E}\left[W(D;h)\theta(Z,X)\right]\right| \\
&\quad - \frac{1}{2}\left|\|\theta\|_{n,2}^2 - \|\theta\|_2^2\right| + \mathbb{E}\left[W(D;h)\theta(Z,X)\right] - \frac{1}{2}\|\theta\|_2^2 \Big\} \\
&\overset{\mathcal{E}}{\gtrsim} \sup_{\theta\in\Theta}\Big\{ -\eta_n\left(\|\theta\|_2 + \eta_n\right) - \frac{1}{4}\left(\|\theta\|_2^2 + \eta_n^2\right) \\
&\quad + \mathbb{E}\left[W(D;h)\theta(Z,X)\right] - \frac{1}{2}\|\theta\|_2^2 \Big\}.
\end{aligned} \tag{E.5}$$

The first inequality holds by the triangle inequality and the second equality holds by the definition of $\mathcal{E}$. Let $\Theta^+(h) = \{\theta \in \Theta : \mathbb{E}\left[W(D;h)\theta(X,Z)\right] > 0\}$. For any $\theta^+ \in \Theta^+(h)$, suppose that $\mathbb{E}\left[W(D;h)\theta^+(X,Z)\right] = \beta\|\theta^+\|_2^2$. By definition of θ^+, we have $\beta > 0$. Let $0 < \kappa \leqslant 1$. Since Θ is star-shaped, we have $\kappa\theta^+ \in \Theta$. By plugging in $\kappa\theta^+$ in (E.5), we have for any $h \in \mathcal{S}(e_n)$ and $\theta^+ \in \Theta^+(h)$ that

$$\mathcal{L}_n(h) \overset{\mathcal{E}}{\gtrsim} \kappa\left(\beta - \frac{3}{4}\kappa\right)\left\|\theta^+\right\|_2^2 - \eta_n\kappa\left\|\theta^+\right\|_2 - \frac{5}{4}\eta_n^2. \tag{E.6}$$

Recall the definition of the solution set $\mathcal{S}(e_n)$. For any $h \in \mathcal{S}(e_n)$, it holds that

$$\mathcal{L}_n(h) \leqslant \inf_{h\in\mathcal{H}} \mathcal{L}_n(h) + e_n \leqslant \mathcal{L}_n(h_\alpha^*) + e_n \overset{\mathcal{E}}{\leqslant} \frac{13}{4}\eta_n^2 + e_n, \tag{E.7}$$

where the second inequality holds by noting that $h_\alpha^* \in \mathcal{H}$ and the last inequality follows from (E.4). Combine (E.6) and (E.7), we get

$$\kappa\left(\beta - \frac{3}{4}\kappa\right)\left\|\theta^+\right\|_2^2 - \eta_n\kappa\left\|\theta^+\right\|_2 - \Delta_n \overset{\mathcal{E}}{\leqslant} 0, \tag{E.8}$$

[1] See the remark under Assumption 4.2.

where we define

$$\Delta_n = \frac{18}{4}\eta_n^2 + e_n. \tag{E.9}$$

Note (E.8) holds for any $0 < \kappa \leqslant 1$. By setting $\kappa = \min\{1, \beta\}$, we have $\beta - 3/4 \cdot \kappa > 0$. Now we solve the quadratic inequality in (E.8), and deduce that on the event $\mathcal{E}$, for all $h \in \mathcal{S}(e_n)$ and $\theta^+ \in \Theta^+(h)$ that

$$\left\|\theta^+\right\|_2 \overset{\mathcal{E}}{\leqslant} \frac{\eta_n\kappa + \sqrt{(\eta_n\kappa)^2 + \kappa\left(4\beta - 3\kappa\right)\Delta_n}}{\kappa\left(2\beta - \frac{3}{2}\kappa\right)}. \tag{E.10}$$

Considering the following two cases.

Case (i) when $\beta \geqslant 1$. If $\beta \geqslant 1$, $\kappa = \min\{\beta, 1\} = 1$. On $\mathcal{E}$, we have

$$\left\|\theta^+\right\|_2^2 \overset{\mathcal{E}}{\leqslant} \left(\frac{\eta_n + \sqrt{\eta_n^2 + (4\beta - 3)\,\Delta_n}}{\left(2\beta - \frac{3}{2}\right)}\right)^2 \leqslant 4\left(\eta_n + \sqrt{\eta_n^2 + \Delta_n}\right)^2 \leqslant 8\left(2\eta_k^2 + \Delta_n\right), \tag{E.11}$$

where the first inequality holds by (E.10), the second inequality holds by noting that $\beta = 1$ maximizes the right-hand side. Plugging this result into (E.8) and rearranging with $\kappa = 1$, we have

$$\begin{aligned}
\mathbb{E}\left[W(D;h)\theta^+(Z)\right] &\leqslant \left\|\theta^+\right\|_2^2 \\
&= \frac{3}{4}\left\|\theta^+\right\|_2^2 + \frac{1}{4}\left\|\theta^+\right\|_2^2 \\
&\overset{\mathcal{E}}{\leqslant} \frac{3}{4}\left\|\theta^+\right\|_2^2 + \eta_n\left\|\theta^+\right\|_2 + \Delta_n \\
&\leqslant 6\left(2\eta_k^2 + \Delta_n\right) + \eta_n\left(4\eta_n + 3\sqrt{\Delta_n}\right) + \Delta_n \\
&\leqslant 16\eta_n^2 + 7\Delta_n + 3\eta_n\sqrt{\Delta_n} \\
&\leqslant 20\eta_n^2 + 11\Delta_n,
\end{aligned} \tag{E.12}$$

where the first inequality holds by noting that $\kappa(\beta - 3/4 \cdot \kappa) = 1/4$ and the third inequality holds by the upper bound of $\left\|\theta^+\right\|_2$ in (E.11).

Case (ii) when $\beta < 1$. If $\beta < 1$, we plug in $\kappa = \beta$. It holds on $\mathcal{E}$ that

$$\beta\left\|\theta^+\right\|_2 \overset{\mathcal{E}}{\leqslant} 2\left(\eta_n + \sqrt{(\eta_n\kappa)^2 + \Delta_n}\right) \leqslant 2\left(2\eta_n + \sqrt{\Delta_n}\right),$$

which suggests that

$$\mathbb{E}\left[W(D;h)\theta^+(X,Z)\right] = \beta\left\|\theta^+\right\|_2^2 \overset{\mathcal{E}}{\leqslant} 2\left(2\eta_n + \sqrt{\Delta_n}\right)\left\|\theta^+\right\|_2. \tag{E.13}$$

Combination of Case (i) and Case (ii). Combining (E.12) and (E.13), we then have for any $h \in \mathcal{S}(e_n)$ and $\theta_k^+ \in \Theta_k^+(h)$ that

$$\begin{aligned}\mathbb{E}\left[W(D;h)\theta^+(X,Z)\right] &\overset{\mathcal{E}}{\leqslant} \max\left\{2\left(\sqrt{\Delta_n}+2\eta_n\right)\|\theta\|_2, 20\eta_n^2+11\Delta_n\right\} \\ &\leqslant \max\left\{C_n\|\theta\|_2, C_n^2\right\},\end{aligned}$$

where $C_n = 5\eta_n + 4\sqrt{\Delta_n}$. We then consider the case in which $\mathbb{E}\left[W(D;h)\theta(X,Z)\right] < 0$ for $\theta \in \Theta\backslash\Theta^+(h)$. Then, for any $h \in \mathcal{S}(e_n)$, $\theta \in \Theta$,

$$\mathbb{E}\left[W(D;h)\theta(X,Z)\right] \overset{\mathcal{E}}{\leqslant} \max\left\{C_n\|\theta\|_2, C_n^2\right\}. \tag{E.14}$$

Now let $\theta_h^*(X,Z) = \arg\min_{\theta\in\Theta}\|\theta - \mathcal{T}h\|_2$. By (E.14), it then holds for any $h \in \mathcal{S}(e_n)$ that

$$\mathbb{E}\left[\mathcal{T}h(X,Z)\theta_h^*(X,Z)\right] \overset{\mathcal{E}}{\leqslant} \max\left\{C_n\|\theta_h^*\|_2, C_n^2\right\}. \tag{E.15}$$

For the left-hand side of (E.15), we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{T}h(X,Z)\theta_h^*(X,Z)\right] &= \mathbb{E}\left[\mathcal{T}h(X,Z)\left(\theta_h^*(X,Z) - \mathcal{T}h(X,Z) + \mathcal{T}h(X,Z)\right)\right] \\ &\geqslant \|\mathcal{T}h\|_2^2 - \|\mathcal{T}h\|_2\|\theta_h^* - \mathcal{T}h\|_2 = \|\mathcal{T}h\|_2^2 + \mathcal{O}(\eta_n),\end{aligned} \tag{E.16}$$

where the first inequality holds by the Cauchy-Schwarz inequality and the last equality holds by Assumption 4.2. For the right-hand side of (E.15), we have,

$$\max\left\{C_n\|\theta_h^*(X,Z)\|_2, C_n^2\right\} \leqslant C_n\left(C_n + \|\theta_h^* - \mathcal{T}h + \mathcal{T}h\|_2\right) \leqslant C_n\left(C_n + \|\mathcal{T}h\|_2\right), \tag{E.17}$$

where the last inequality holds by the triangle inequality and Assumption 4.2. Combining (E.16) and (E.17) with (E.15), for all $h \in \mathcal{S}(e_n)$, we have:

$$\|\mathcal{T}h\|_2^2 - C_n^2 - C_n\|\mathcal{T}h\|_2 \overset{\mathcal{E}}{\leqslant} 0,$$

which implies that

$$\begin{aligned}\|\mathcal{T}h\|_2 &\overset{\mathcal{E}}{\leqslant} \frac{1}{2}\left(C_n + \sqrt{C_n^2 + 4C_n^2}\right) = \mathcal{O}(C_n) \\ &= \mathcal{O}\left(5\eta_n + 4\sqrt{18/4\cdot\eta_n^2 + e_n}\right) \\ &= \mathcal{O}\left(\sqrt{e_n}\right) + \mathcal{O}\left(\eta_n\right),\end{aligned}$$

where the third equality holds by the definition of C_n and the definition of Δ_n in (E.9). □

E.2 Proof of Theorem 4.8

We prove that, on event $\mathcal{E}$, any $h \in \mathcal{S}(e_n)$ converges in probability to h_α^* with respect to $\|\cdot\|_\infty$ and hence $\|\cdot\|_2$, when $e_n = \mathcal{O}(\eta_n^2)$.

Proof. We first show that it converges in probability in $\|\cdot\|_\infty$ by using two lemmas that investigate the property of the loss function $\mathcal{L}_n$. Lemma E.1 below gives a lower bound of the empirical loss on event $\mathcal{E}$. Lemma E.2 shows that the RMSE $\|\mathcal{T}h\|_2^2$ is continuous on $(\mathcal{H}, \|\cdot\|_\infty)$.

Lemma E.1 (Lower Bound of Empirical Loss)**.** Suppose that Assumptions 4.3 and 4.2 hold. On the event $\mathcal{E}$, for all $h \in \mathcal{H}$, we have

$$\mathcal{L}_n(h) = \sup_{\theta \in \Theta} \mathbb{E}_n \left[W(D;h)\theta(X,Z) \right] - \frac{1}{2}\|\theta\|_{n,2}^2 \geqslant \min\left\{ \frac{\eta_n}{2}\|\mathcal{T}h\|_2 \|\mathcal{T}h\|_2^2 \right\} - \mathcal{O}\left(\eta_n^2\right).$$

Proof. See §G for a detailed proof. □

Lemma E.2 (Continuity of Linear Operator)**.** Suppose the Assumption 4.7 holds. $\|\mathcal{T}h\|_2^2$ is continuous on $(\mathcal{H}, \|\cdot\|_\infty)$.

Proof. See §G for a detailed proof. □

We are ready to prove consistency. For any $\epsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left[\|h - h_\alpha^*\|_\infty > \epsilon, h \in \mathcal{S}(e_n)\right] \\
&\quad \leqslant \mathbb{P}\left[\inf_{h \in \mathcal{H}, \|h - h_\alpha^*\|_\infty > \epsilon} \mathcal{L}_n(h) \leqslant \mathcal{O}(\eta_n^2) \right] \\
&\quad \leqslant \mathbb{P}\left[\min\left\{ \inf_{h \in \mathcal{H}, \|h - h_\alpha^*\|_\infty > \epsilon} \frac{\eta_n}{2}\|\mathcal{T}h\|_2, \inf_{h \in \mathcal{H}, \|h - h_\alpha^*\|_\infty > \epsilon} \|\mathcal{T}h\|_2^2 \right\} \leqslant \mathcal{O}\left(\eta_n^2\right) \right],
\end{aligned}$$

where the second inequality holds by Lemma E.1. Denote $\varphi_\epsilon := \inf_{h \in \mathcal{H}, \|h - h_\alpha^*\|_\infty > \epsilon} \|\mathcal{T}h\|_2$. Assumption 4.1, 4.3, and Lemma E.2 implies that $\mathcal{H}$ is compact and $\|\mathcal{T}h\|_2$ is continuous on $(\mathcal{H}, \|\cdot\|_\infty)$. Hence φ_ϵ is strictly positive. We then have

$$\mathbb{P}\left[\|h - h_\alpha^*\|_\infty > \epsilon, h \in \mathcal{S}(e_n)\right] \leqslant \mathbb{P}\left[\min\left\{ \frac{\eta_n}{2}\varphi_\epsilon, \varphi_\epsilon^2 \right\} \leqslant \mathcal{O}\left(\eta_n^2\right) \right],$$

which converges to zero as n goes to infinity. Since the supremum norm is stronger than the L_2 norm under a finite measure:

$$\mathbb{P}\left[\|h - h_\alpha^*\|_2 > \epsilon, h \in \mathcal{S}(e_n)\right] \leqslant \mathbb{P}\left[\|h - h_\alpha^*\|_\infty > \epsilon, h \in \mathcal{S}(e_n)\right],$$

we have $\mathbb{P}\left[\|h - h_\alpha^*\|_2 > \epsilon, h \in \mathcal{S}(e_n)\right]$ converges to 0 as n goes to infinity. That is,

$$\lim_{n \to \infty} \mathbb{P}\left[\|h - h_\alpha^*\|_2 > \epsilon, h \in \mathcal{S}(e_n)\right] = 0.$$

Thus, we conclude the proof. □

E.3 Proof of Theorem 4.14

We prove that if we set $\lambda_n \geqslant \eta_n^{-(1+\epsilon_R)}$ where ϵ_R is an arbitrary positive real number, then on the event $\mathcal{E}$, $\widehat{h}_R^\pi$ converges in probability to h_α^* with respect to $\|\cdot\|_\infty$ and hence $\|\cdot\|_2$.

Proof. We use a similar argument as in the proof of Theorem 4.8. But we first need to quantify an upper bound of the empirical loss of the regularized estimator:

Lemma E.3 (Upper Bound of the Empirical Loss of Regularized Estimator)**.** Suppose that Assumption 4.1, Assumption 4.3, and Assumption 4.13 hold. If we set $\lambda_n \geqslant \eta_n^{-(1+\epsilon_R)}$ where $0 < \epsilon_R < 1$ is arbitrary, then on the event $\mathcal{E}$, $\mathcal{E}_n(\widehat{h}_R^\pi) = \mathcal{O}\left(\eta_n^{1+\epsilon_R}\right)$ and $\mathcal{L}_n(\widehat{h}_R^\pi) = \mathcal{O}\left(\eta_n^{1+\epsilon_R}\right)$.

Proof. See §G for a detailed proof. □

Therefore, for any $\epsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left[\left\|\widehat{h}_R^\pi - h_\alpha^*\right\|_\infty > \epsilon\right] \\
&\quad\leqslant \mathbb{P}\left[\inf_{h\in\mathcal{H}, \left\|h-h_\alpha^*\right\|_\infty > \epsilon} \mathcal{L}_n(h) \leqslant \mathcal{O}\left(\eta_n^{1+\epsilon_R}\right)\right] \\
&\quad\leqslant \mathbb{P}\left[\min\left\{\inf_{h\in\mathcal{H}, \left\|h-h_\alpha^*\right\|_\infty > \epsilon} \frac{\eta_n}{2}\left\|\mathcal{T}h\right\|_2, \inf_{h\in\mathcal{H}, \left\|h-h_\alpha^*\right\|_\infty > \epsilon} \left\|\mathcal{T}h\right\|_2^2\right\} \leqslant \mathcal{O}\left(\eta_n^{1+\epsilon_R}\right)\right],
\end{aligned}$$

where the first inequality holds by Lemma E.3, and the second inequality holds by Lemma E.1. Denote $\varphi_\epsilon := \inf_{h\in\mathcal{H}, \left\|h-h_\alpha^*\right\|_\infty > \epsilon} \left\|\mathcal{T}h\right\|_2$. Assumption 4.1, 4.3, and Lemma E.2 implies that $\mathcal{H}$ is compact and $\left\|\mathcal{T}h\right\|_2$ is continuous on $(\mathcal{H}, \|\cdot\|_\infty)$. Hence φ_ϵ is strictly positive. We then have

$$\mathbb{P}\left[\left\|\widehat{h}_R^\pi - h_\alpha^*\right\|_\infty > \epsilon\right] \leqslant \mathbb{P}\left[\min\left\{\frac{\eta_n}{2}\varphi_\epsilon, \varphi_\epsilon^2\right\} \leqslant \mathcal{O}\left(\eta_n^{1+\epsilon_R}\right)\right],$$

which converges to zero as n goes to infinity. Since the supremum norm is stronger than the L_2 norm under a finite measure:

$$\mathbb{P}\left[\left\|\widehat{h}_R^\pi - h_\alpha^*\right\|_2 > \epsilon\right] \leqslant \mathbb{P}\left[\left\|\widehat{h}_R^\pi - h_\alpha^*\right\|_\infty > \epsilon\right],$$

we have $\mathbb{P}\left[\left\|\widehat{h}_R^\pi - h_\alpha^*\right\|_2 > \epsilon\right]$ converges to 0 as n goes to infinity:

$$\lim_{n\to\infty} \mathbb{P}\left[\left\|\widehat{h}_R^\pi - h_\alpha^*\right\|_2 > \epsilon\right] = 0.$$

Thus, we conclude the proof. □

E.4 Decomposition of the Regret with Pessimism & Proof of Corollary 4.6

In this section, we study the regret of the estimated policy $\widehat{\pi}$ with pessimism. The result in this section will be utilized in §E.5 to derive the regret. Recall the definition of $\widehat{h}^{\pi}$ and $\widehat{\pi}$:

$$\widehat{h}^{\pi} := \arg \inf_{h \in \mathcal{S}(e_n)} v(h, \pi), \widehat{\pi} := \arg \sup_{\pi} \inf_{h \in \mathcal{S}(e_n)} v(h, \pi) = \arg \sup_{\pi} v(\widehat{h}^{\pi}, \pi). \tag{E.18}$$

The regret of policy $\widehat{\pi}$ is bounded by

$$\begin{aligned} \text{Regret}(\widehat{\pi}) &= v_{\alpha}^{\pi^*} - v_{\alpha}^{\widehat{\pi}} \\ &= \underbrace{v_{\alpha}^{\pi^*} - v(\widehat{h}^{\pi^*}, \pi^*)}_{\text{(i)}} + \underbrace{v(\widehat{h}^{\pi^*}, \pi^*) - v(\widehat{h}^{\widehat{\pi}}, \widehat{\pi})}_{\text{(ii)}} + \underbrace{v(\widehat{h}^{\widehat{\pi}}, \widehat{\pi}) - v_{\alpha}^{\widehat{\pi}}}_{\text{(iii)}}, \end{aligned} \tag{E.19}$$

Here, (ii) $\leqslant 0$ holds by the optimality of $\widehat{\pi}$ and (iii) $\overset{\mathcal{E}}{\leqslant} 0$ holds by the definition of $\widehat{h}^{\widehat{\pi}}$ in (E.18) and the fact that $h_{\alpha}^* \in \mathcal{S}(e_n)$ on event $\mathcal{E}$ by Theorem 4.5. Therefore, we just need to bound (i). This proves Corollary 4.6.

We now establish an upper bound for (i) as follows.

E.5 Proof of Theorem 4.11

We upper bound the regret of the solution set algorithm in the context of IV. By Corollary 4.6, we can focus on the estimation error of the average reward function for the optimal interventional policy π^*, i.e., Term (i) in (E.19).

Proof. By definition of the change of measure function (Assumption 4.10), we have:

$$\begin{aligned} v_{\alpha}^{\pi^*} - v(\widehat{h}^{\pi^*}, \pi^*) &= \mathbb{E}_{p_{\text{in}}^{\pi^*}} \left[h_{\alpha}^*(A, X) - \widehat{h}^{\pi^*}(A, X) \right] \\ &= \mathbb{E} \left[(h_{\alpha}^*(A, X) - \widehat{h}^{\pi^*}(A, X)) \cdot \mathbb{E} \left[b(X, Z) p_{Y \mid A, X, Z}(h_{\alpha}^*(A, X)) \mid A, X \right] \right]. \end{aligned}$$

Then by the tower property,

$$\begin{aligned} v_{\alpha}^{\pi^*} - v(\widehat{h}^{\pi^*}, \pi^*) &= \mathbb{E} \left[\mathbb{E} \left[(h_{\alpha}^*(A, X) - \widehat{h}^{\pi^*}(A, X)) b(X, Z) p_{Y \mid A, X, Z}(h_{\alpha}^*(A, X)) \mid A, X \right] \right] \\ &= \mathbb{E}[p_{Y \mid A, X, Z}(h_{\alpha}^*(A, X))(h_{\alpha}^*(A, X) - \widehat{h}^{\pi^*}(A, X)) b(X, Z)] \\ &= \mathbb{E} \left[\mathbb{E} \left[p_{Y \mid A, X, Z}(h_{\alpha}^*(A, X))(h_{\alpha}^*(A, X) - \widehat{h}^{\pi^*}(A, X)) \mid X, Z \right] b(X, Z) \right] \\ &\leqslant \| \widehat{h}^{\pi^*} - h_{\alpha}^* \|_{\text{ps}} \cdot \| b \|_2 , \end{aligned}$$

where the last inequality holds by the Cauchy-Schwarz inequality. Denote $\mathcal{E}_{\epsilon} := \mathcal{E} \cap \left\{ \widehat{h}^{\pi^*} \in \mathcal{H}_{\epsilon} \right\}$. Thus we can bound the regret as

$$\text{Regret}(\widehat{\pi}) \overset{\mathcal{E}_{\epsilon}}{\leqslant} \| \widehat{h}^{\pi^*} - h_{\alpha}^* \|_{\text{ps}} \cdot \| b \|_2 \leqslant c_0 \cdot \left\| \mathcal{T} \widehat{h}^{\pi^*} \right\|_2 \cdot \| b \|_2 = c_0 \cdot \| b \|_2 \cdot [\mathcal{O}(\sqrt{e_n}) + \mathcal{O}(\eta_n)],$$

where the first inequality holds by the regret decomposition in §E.4, the second inequality holds by Assumption 4.9, and the last inequality holds by the Theorem 4.5. Hence, we complete the proof of Theorem 4.11. □

E.6 Proof of Theorem 4.15 and Lemma 4.12

In this section, we prove that if the regularized parameter λ_n is set to $\lambda_n = \eta_n^{-(1+\epsilon_R)}$, then on the event $\mathcal{E}$,the regret of the regularized algorithm is of order $\eta_n^{1-\epsilon_R}$.

Proof. We first decompose the regret corresponding to $\widehat{\pi}_R$ as:

$$\begin{aligned}
\text{Regret}(\widehat{\pi}_R) &= v_\alpha^{\pi^*} - v_\alpha^{\widehat{\pi}_R} \\
&= \left[v_\alpha^{\pi^*} + \lambda_n \mathcal{E}_n\left(h_\alpha^*\right)\right] - \left[v(\widehat{h}_R^{\widehat{\pi}_R}, \widehat{\pi}_R) + \lambda_n \mathcal{E}_n(\widehat{h}_R^{\widehat{\pi}_R})\right] + \left[v(\widehat{h}_R^{\widehat{\pi}_R}, \widehat{\pi}_R) + \lambda_n \mathcal{E}_n(\widehat{h}_R^{\widehat{\pi}_R})\right] \\
&\quad - \left[v\left(h_\alpha^*, \widehat{\pi}_R\right) + \lambda_n \mathcal{E}_n\left(h_\alpha^*\right)\right].
\end{aligned}$$

By the optimality of $\widehat{h}_R^{\widehat{\pi}}$, we have

$$\begin{aligned}
\text{Regret}(\widehat{\pi}_R) &\leqslant \left[v_\alpha^{\pi^*} + \lambda_n \mathcal{E}_n\left(h_\alpha^*\right)\right] - \left[v(\widehat{h}_R^{\widehat{\pi}_R}, \widehat{\pi}_R) + \lambda_n \mathcal{E}_n(\widehat{h}_R^{\widehat{\pi}_R})\right] \\
&\leqslant \left[v_\alpha^{\pi^*} + \lambda_n \mathcal{E}_n\left(h_\alpha^*\right)\right] - \sup_{\pi} \inf_{h \in \mathcal{H}} \{v\left(h, \pi\right) + \lambda_n \mathcal{E}_n(h)\} \\
&\leqslant \left[v_\alpha^{\pi^*} + \lambda_n \mathcal{E}_n(h_\alpha^*)\right] - \left[v(\widehat{h}_R^{\pi^*}, \pi^*) + \lambda_n \mathcal{E}_n(\widehat{h}_R^{\pi^*})\right].
\end{aligned}$$

After rearranging the terms, we get

$$\begin{aligned}
\text{Regret}(\widehat{\pi}_R) &= \mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}_R^{\pi^*}(A, X)\right] + \lambda_n \mathcal{E}_n(h_\alpha^*) - \lambda_n \mathcal{E}_n(\widehat{h}_R^{\pi^*}) \\
&\leqslant \mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}_R^{\pi^*}(A, X)\right] + \lambda_n \mathcal{E}_n\left(h_\alpha^*\right). \qquad \text{(E.20)}
\end{aligned}$$

This concludes the proof of Lemma 4.12.

Following the same argument as in Section E.5, we conclude that for the first term,

$$\mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}_R^{\pi^*}(A, X)\right] \overset{\mathcal{E}_{R\epsilon}}{\leqslant} c_0 \left\|b\right\|_2 \cdot \left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2. \qquad \text{(E.21)}$$

For the second term, we have by Theorem 4.5 (ii) that

$$\lambda \mathcal{E}_n(h_\alpha^*) \leqslant \lambda \mathcal{L}_n(h_\alpha^*) \leqslant \mathcal{O}(\eta_n^{1-\epsilon_R}). \qquad \text{(E.22)}$$

Moreover, using the fact that $\text{Regret}(\widehat{\pi}_R) \geqslant 0$ and rearranging terms, we have

$$\begin{aligned}
\mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}_R^{\pi^*}(A, X)\right] &\geqslant \lambda_n \mathcal{E}_n(\widehat{h}_R^{\pi^*}) - \lambda_n \mathcal{E}_n\left(h_\alpha^*\right) \\
&\overset{\mathcal{E}_{R\epsilon}}{\gtrsim} \lambda_n \mathcal{L}_n(\widehat{h}_R^{\pi^*}) - \lambda_n \mathcal{O}(\eta_n^2) \\
&\geqslant \lambda_n \min\left\{\frac{\eta_n}{2}\left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2, \left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2^2\right\} - \lambda_n \mathcal{O}(\eta_n^2). \qquad \text{(E.23)}
\end{aligned}$$

We now show that $\left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2 = \mathcal{O}(\eta_n)$. When $\eta_n/2 \cdot \left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2 \geqslant \left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2^2$, the result is immediate. When $\eta_n/2 \cdot \left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2 < \left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2^2$, combining (E.21) and (E.23), we have

$$c_0 \|b\|_2 \cdot \left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2 \overset{\mathcal{E}_{R\epsilon}}{\gtrsim} \lambda_n \cdot \frac{\eta_n}{2} \left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2 - \lambda_n \mathcal{O}(\eta_n^2).$$

Solving for $\left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2$ and substituting $\lambda_n = \eta_n^{-(1+\epsilon_R)}$ yields

$$\left\|\mathcal{T}\widehat{h}_R^{\pi^*}\right\|_2 = \mathcal{O}(\eta_n).$$

Substituting this into (E.21), we know the first term of (E.20) is upper bounded by $\mathcal{O}(\eta_n)$. By (E.22), we know the second term of (E.20) is upper bounded by $\mathcal{O}(\eta_n^{1-\epsilon_R})$. By summing up the two terms, we conclude that conditioning on the event $\mathcal{E}_{R\epsilon}$,

$$\text{Regret}(\widehat{\pi}_R) = \mathcal{O}\left(\eta_n^{1-\epsilon_R}\right).$$

Therefore, we complete the proof. □

E.7 Proof of Corollary B.4

We present the regret for minimizing the quantile-based risk measures. The idea is to aggregate the individual regret for each α_i.

Proof. For each $i \in [m]$, we define an event $\mathcal{E}_i$ as

$$\begin{aligned}\mathcal{E}_i := \Big\{ &\left|\mathbb{E}_n\left[W_i(D;h_{\alpha_i})\theta(X,Z)\right] - \mathbb{E}\left[W_i(D;h_{\alpha_i})\theta(X,Z)\right]\right| \leqslant \eta_n\left(\|\theta\|_2 + \eta_n\right), \\ &\left|\|\theta\|_{n,2}^2 - \|\theta\|_2^2\right| \leqslant \frac{1}{2}\left(\|\theta\|_2^2 + \eta_n^2\right), \forall h_{\alpha_i} \in \mathcal{H}, \forall \theta \in \Theta\Big\},\end{aligned} \tag{E.24}$$

and then define $\mathcal{E}' = \bigcap_{i=1}^m \mathcal{E}_i$. So $\mathcal{E}'$ holds with probability at least $1 - 2m\xi$. We can upper bound the regret of $\widehat{\pi}^{\text{RM}}$ as:

$$\begin{aligned}\text{Regret}(\widehat{\pi}^{\text{RM}}) &= \mathbb{E}_{p_{\text{in}}^{\pi^*\text{RM}}}\left[\sum_{i=1}^m \phi(\alpha_i) h_{\alpha_i}^*(A,X)\right] - \mathbb{E}_{p_{\text{in}}\widehat{\pi}^{\text{RM}}}\left[\sum_{i=1}^m \phi(\alpha_i) h_{\alpha_i}(A,X)\right] \\ &\overset{\mathcal{E}'}{\leqslant} \mathbb{E}_{p_{\text{in}}^{\pi^*\text{RM}}}\left[\sum_{i=1}^m \phi(\alpha_i)(h_{\alpha_i}^*(A,X) - \widehat{h}_{\alpha_i}^{\pi^*}(A,X))\right] \\ &\leqslant M \sum_{i=1}^m \mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_{\alpha_i}^*(A,X) - \widehat{h}_{\alpha_i}^{\pi^*}(A,X))\right].\end{aligned}$$

For the regularized algorithm, we have a similar upper bound:

$$\begin{aligned}\text{Regret}\left(\widehat{\pi}_R^{\text{RM}}\right) &\leqslant \mathbb{E}_{p_{\text{in}}^{\pi^*\text{RM}}}\left[\sum_{i=1}^m \phi(\alpha_i)(h_{\alpha_i}^*(A,X) - \widehat{h}_{R\alpha_i}^{\pi^*}(A,X))\right] + \lambda_n(\mathcal{E}_n(h^*) - \mathcal{E}_n(\widehat{h}_R^{\pi^*}) \\ &\leqslant M \sum_{i=1}^m \mathbb{E}_{p_{\text{in}}^{\pi}}\left[h_{\alpha_i}^*(A,X) - \widehat{h}_{R\alpha_i}^{\pi^*}(A,X)\right] + \lambda_n(\mathcal{E}_n(h^*) - \mathcal{E}_n(\widehat{h}_R^{\pi^*}).\end{aligned}$$

The rest of the proof now follows from applying the proof in Theorem 4.11 and Theorem 4.15 to each i. □

F Proof of Main Results of §C

In this section, we give the proofs of the main theorems in §C. The goal is to establish the regret for the policy learning algorithms in the case that we observe NC in ODCP.

F.1 Proof of Theorem C.2

We show that given Assumption C.1, h_1^* and h_2^* satisfy the conditional moment restrictions given in Theorem C.2.

Proof. From (C.1), we have

$$\begin{aligned}
0 &= \mathbb{E}\left[W(D;h_1^*) - h_2^*(V,A,X) \,|\, E=e, A=a, X=x\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[W(D;h_1^*) - h_2^*(V,A,X) \,|\, U,E,A,X\right] \,|\, E=e, A=a, X=x\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[W(D;h_1^*) - h_2^*(V,A,X) \,|\, U,A,X\right] \,|\, E=e, A=a, X=x\right],
\end{aligned}$$

where the last equality holds by the Assumption C.1 (i). Now by Assumption C.1 (ii), we have

$$\mathbb{E}\left[W(D;h_1^*) \,|\, U,A,X\right] = \mathbb{E}\left[h_2^*(V,A,X) \,|\, U,A,X\right]. \tag{F.1}$$

From (C.2), $\forall a' \in \mathcal{A}$, $\forall a \in \mathcal{A}$, we have

$$\begin{aligned}
0 &= \mathbb{E}\left[h_2^*(V,a',X) \,|\, X=x\right] = \mathbb{E}\left[\mathbb{E}\left[h_2^*(V,a',X) \,|\, U,X\right] \,|\, X=x\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[h_2^*(V,a',X) \,|\, U,A=a,X\right] \,|\, X=x\right],
\end{aligned} \tag{F.2}$$

where the third equality holds by Assumption C.1 (i). Now by combining (F.1) and (F.2), we have

$$0 = \mathbb{E}\left[\mathbb{E}\left[W(D;h_1^*) \,|\, U,A=a,X\right] \,|\, X=x\right] = \mathbb{E}\left[W(D;h_1^*) \,|\, do(A=a), X=x\right].$$

By the definition of $W(D;h_1^*)$, we conclude that

$$\mathbb{P}\left[Y \leqslant h_1^* \,|\, X=x, \mathrm{do}(A=a)\right] = \alpha.$$

Therefore, we complete the proof. □

F.2 Proof of Theorem D.9

In this section, we analyze the regret of the solution set algorithm for the negative controls.

Proof. We first note that the regret decomposition for the case of negative control is exactly the same as that of the instrumental variable. Therefore, we begin by considering the term (i) in (E.19). By the change of measure Assumption C.3, we have:

$$\begin{aligned} v_\alpha^{\pi^*} - v(\widehat{h}_1^{\pi^*}, \pi^*) &= \mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A,X) - \widehat{h}_1^{\pi^*}(A,X)\right] \\ &= \mathbb{E}\left[(h_\alpha^*(A,X) - \widehat{h}^{\pi^*}(A,X)_1)\mathbb{E}\left[b_1(E,A,X)p_{y\,|\,E,A,X}(h_\alpha^*(A,X))\,|\,A,X\right]\right]. \end{aligned}$$

Then by Tower property,

$$\begin{aligned} v_\alpha^{\pi^*} - v(\widehat{h}_1^{\pi^*}, \pi^*) &= \mathbb{E}\left[\mathbb{E}\left[(h_\alpha^*(A,X) - \widehat{h}_1^{\pi^*}(A,X))b_1(E,A,X)p_{y\,|\,E,A,X}(h_\alpha^*(A,X))\,|\,A,X\right]\right] \\ &= \mathbb{E}[p_{y\,|\,E,A,X}(h_\alpha^*(A,X))(h_\alpha^*(A,X) - \widehat{h}_1^{\pi^*}(A,X))b_1(E,A,X)] \\ &= \mathbb{E}\left[\mathbb{E}\left[p_{y\,|\,E,A,X}(h_\alpha^*(A,X))(h_\alpha^*(A,X) - \widehat{h}_1^{\pi^*}(A,X))\,|\,E,A,X\right]b_1(E,A,X)\right]. \end{aligned}$$

Now let $\widetilde{\mathcal{E}}_\epsilon := \widetilde{\mathcal{E}} \cap \left\{\widehat{h}_1^{\pi^*} \in \mathcal{H}_{1\epsilon}\right\}$. By Cauchy-Schwarz and the form of directional derivative in Lemma G.1 (with a slight change of the conditional variables), we have

$$\begin{aligned} \text{Regret}(\widehat{\pi}) &\overset{\widetilde{\mathcal{E}}_\epsilon}{\leqslant} \|\widehat{h}_1^{\pi^*} - h_\alpha^*\|_{\text{ps}} \cdot \|b_1\|_2 \\ &\leqslant c_1 \|b_1\|_2 \cdot \left\|\mathbb{E}[W(D;\widehat{h}_1^{\pi^*})\,|\,E,A,X]\right\|_2, \end{aligned} \tag{F.3}$$

where the last inequality holds by Assumption D.8. We now obtain an upper bound for $\left\|\mathbb{E}[W\left(D;\widehat{h}_1^{\pi^*}\right)\,|\,E,A,X]\right\|_2$:

$$\begin{aligned} \left\|\mathbb{E}[W\left(D;\widehat{h}^{\pi^*}\right)\,|\,E,A,X]\right\|_2 &= \left\|\mathcal{T}_1\widehat{h}^{\pi^*} + \mathbb{E}\left[\widehat{h}_2^{\pi^*}(W,A,X)\,|\,E,A,X\right]\right\|_2 \\ &\leqslant \left\|\mathcal{T}_1\widehat{h}^{\pi^*}\right\|_2 + \left\|\mathbb{E}\left[\widehat{h}_2^{\pi^*}(W,A,X)\,|\,E,A,X\right]\right\|_2 \\ &= \left\|\mathcal{T}_1\widehat{h}^{\pi^*}\right\|_2 + \left\|\int_{\mathcal{W}} \widehat{h}_2^{\pi^*}(w,A,X)p(w\,|\,E,A,X)\mathrm{d}w\right\|_2. \end{aligned}$$

By Assumption C.3 (ii), we have

$$\begin{aligned} \left\|\mathbb{E}[W\left(D;\widehat{h}^{\pi^*}\right)\,|\,E,A,X]\right\|_2 &\leqslant \left\|\mathcal{T}_1\widehat{h}^{\pi^*}\right\|_2 + \left\|\int_{\mathcal{W}} \widehat{h}_2^{\pi^*}(w,A',X)p(w\,|\,X)b_2(w,E,A,X,A')\mathrm{d}w\right\|_2 \\ &\leqslant \left\|\mathcal{T}_1\widehat{h}^{\pi^*}\right\|_2 + \|b_2\|_\infty \left\|\int_{\mathcal{W}} \widehat{h}_2^{\pi^*}(w,A',X)p_{\text{ob}}(w\,|\,A',X)\mathrm{d}w\right\|_2 \\ &= \left\|\mathcal{T}_1\widehat{h}^{\pi^*}\right\|_2 + \|b_2\|_\infty \left\|\mathcal{T}_2\widehat{h}^{\pi^*}\right\|_2. \end{aligned}$$

Substitute the above result back to (F.3), we complete the proof:

$$\begin{aligned}\text{Regret}(\widehat{\pi}) &\overset{\widetilde{\mathcal{E}}_\epsilon}{\leqslant} c_1 \left\|b_1\right\|_2 \cdot \left\|\mathcal{T}_1 \widehat{h}^{\pi^*}\right\|_2 + c_1 \left\|b_1\right\|_2 \cdot \left\|b_2\right\|_\infty \left\|\mathcal{T}_2 \widehat{h}^{\pi^*}\right\|_2 \\ &= c_1(1+\left\|b_2\right\|_\infty) \left\|b_1\right\|_2 \cdot \left(\mathcal{O}\left(\sqrt{e_n}\right) + \mathcal{O}\left(\eta_n\right)\right),\end{aligned}$$

where the equality holds by Theorem D.5 (ii). □

F.3 Proof of Theorem D.11

We show that our estimated structural quantile function $\widehat{h}^\pi_{R1}$ is consistent to h^*_α in the ℓ_2-norm. The proof is similar to the proof of Theorem 4.14, except that the dimension of $\widehat{h}^\pi_R$ is now two. We deduce the result by first showing that $\widehat{h}^\pi_R = (\widehat{h}^\pi_{R1}, \widehat{h}^\pi_{R2})$ is consistent to $h^* = (h^*_\alpha, {h_2}^*)$ in $\|\cdot\|_{\sup}$.

Proof. For any $\epsilon > 0$, we have

$$\begin{aligned}\mathbb{P}\left[\left\|\widehat{h}^\pi_R - h^*\right\|_{\sup} > \epsilon\right] &\leqslant \mathbb{P}\left[\inf_{h\in\mathcal{H}, \|h-h^*\|_{\sup}>\epsilon} \mathcal{L}_{1,n}(h) \leqslant \mathcal{O}(\eta_n^{1+\epsilon_R})\right] \\ &\leqslant \mathbb{P}\left[\min\left\{\inf_{h\in\mathcal{H}, \|h-h^*\|_{\sup}>\epsilon} \frac{\eta_n}{2}\left\|\mathcal{T}_1 h\right\|_2, \inf_{h\in\mathcal{H}, \|h-h^*\|_{\sup}>\epsilon} \left\|\mathcal{T}_1 h\right\|_2^2\right\} \leqslant \mathcal{O}\left(\eta_n^{1+\epsilon_R}\right)\right].\end{aligned}$$

Denote $\varphi_\epsilon := \inf_{h\in\mathcal{H}, \|h-h^*\|_\infty>\epsilon} \left\|\mathcal{T}_1 h\right\|_2$. Since Lemma H.3 states that $\left\|\mathcal{T}_1 h\right\|_2$ is continuous on $(\mathcal{H}, \|\cdot\|_{\sup})$, φ_ϵ is strictly positive. We then have

$$\mathbb{P}\left[\|\widehat{h}^\pi_R - h^*\|_{\sup} > \epsilon\right] \leqslant \mathbb{P}\left[\min\left\{\frac{\eta_n}{2}\varphi_\epsilon, \varphi_\epsilon^2\right\} \leqslant \mathcal{O}(\eta_n^{1+\epsilon_R})\right],$$

which converges to 0 as n goes to infinity. The result follows from the fact that the supremum norm is stronger than the L_2 norm under a finite measure:

$$\mathbb{P}\left[\left\|\widehat{h}^\pi_R - h^*\right\|_2 > \epsilon\right] \leqslant \mathbb{P}\left[\left\|\widehat{h}^\pi_R - h^*\right\|_\infty > \epsilon\right],$$

which converges to 0 as n goes to infinity. Hence, we have $\|\widehat{h}^\pi_{R1} - h^*_\alpha\|_\infty = o_p(1)$ and $\|\widehat{h}^\pi_{R1} - h^*_\alpha\|_2 = o_p(1)$. □

F.4 Proof of Theorem D.12

In this section, we analyze the regret of the regularized algorithm for NC. The proof is similar to the proof of Theorem 4.15.

Proof. We first decompose the regret corresponding to $\widehat{\pi}_R$ as:

$$\begin{aligned}\text{Regret}(\widehat{\pi}_R) &= v_\alpha^{\pi^*} - v_\alpha^{\widehat{\pi}_R} \\ &= \left[v_\alpha^{\pi^*} + \lambda_n \mathcal{E}_n\left(h^*\right)\right] - \left[v(\widehat{h}^{\widehat{\pi}_R}_{R1}, \widehat{\pi}_R) + \lambda_n \mathcal{E}_n(\widehat{h}^{\widehat{\pi}_R}_R)\right] + \left[v(\widehat{h}^{\widehat{\pi}_R}_{R1}, \widehat{\pi}_R) + \lambda_n \mathcal{E}_n(\widehat{h}^{\widehat{\pi}_R}_R)\right] \\ &\quad - \left[v\left(h^*_\alpha, \widehat{\pi}_R\right) + \lambda_n \mathcal{E}_n\left(h^*\right)\right].\end{aligned}$$

By the optimality of $\widehat{h}_R^{\widehat{\pi}}$, we can drop the last two terms. It follows that

$$\begin{aligned}\text{Regret}(\widehat{\pi}_R) &\leqslant \left[v_\alpha^{\pi^*} + \lambda_n \mathcal{E}_n(h^*)\right] - \left[v(\widehat{h}_{R1}^{\widehat{\pi}_R}, \widehat{\pi}_R) + \lambda_n \mathcal{E}_n(\widehat{h}_R^{\widehat{\pi}_R})\right] \\ &\leqslant \left[v_\alpha^{\pi^*} + \lambda_n \mathcal{E}_n(h^*)\right] - \sup_{\pi} \inf_{h \in \mathcal{H}} \{v(h_1, \pi) + \lambda_n \mathcal{E}_n(h)\} \\ &\leqslant \left[v_\alpha^{\pi^*} + \lambda_n \mathcal{E}_n(h^*)\right] - \left[v(\widehat{h}_{R1}^{\pi^*}, \pi^*) + \lambda_n \mathcal{E}_n(\widehat{h}_R^{\pi^*})\right].\end{aligned}$$

Then by rearranging terms, we have

$$\begin{aligned}\text{Regret}(\widehat{\pi}_R) &= \mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}_{R1}^{\pi^*}(A, X)\right] + \lambda_n \mathcal{E}_n(h^*) - \lambda_n \mathcal{E}_n(\widehat{h}_R^{\pi^*}) \\ &\leqslant \mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}_{R1}^{\pi^*}(A, X)\right] + \lambda_n \mathcal{E}_n(h^*). \end{aligned} \tag{F.4}$$

Following the same argument as in the proof of Theorem D.9, we conclude that for the first term,

$$\mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}_{R1}^{\pi^*}(A, X)\right] \overset{\widetilde{\mathcal{E}}_{R\epsilon}}{\leqslant} c_1 \|b_1\|_2 \cdot \left\|\mathcal{T}_1 \widehat{h}^{\pi^*}\right\|_2 + c_1 \|b_1\|_2 \cdot \|b_2\|_\infty \left\|\mathcal{T}_2 \widehat{h}^{\pi^*}\right\|_2. \tag{F.5}$$

For the second term, we have by Theorem D.5 (ii) that

$$\lambda_n \mathcal{E}_n(h^*) \leqslant \lambda_n \mathcal{L}_n(h^*) \leqslant \mathcal{O}(\eta_n^{1-\epsilon_R}). \tag{F.6}$$

Moreover, using the fact that $\text{Regret}(\widehat{\pi}_R) \geqslant 0$ and rearranging terms, we have

$$\mathbb{E}_{p_{\text{in}}^{\pi^*}}\left[h_\alpha^*(A, X) - \widehat{h}_{R1}^{\pi^*}(A, X)\right] \geqslant \lambda_n \mathcal{E}_n(\widehat{h}_R^{\pi^*}) - \lambda_n \mathcal{E}_n(h^*) \overset{\widetilde{\mathcal{E}}_\epsilon}{\geqslant} \lambda_n \mathcal{L}_n(\widehat{h}_R^{\pi^*}) - \lambda_n \mathcal{O}(\eta_n^2).$$

We can further lower bound it as

$$\lambda_n \min\left\{\frac{\eta_n}{2}\left\|\mathcal{T}_1 \widehat{h}_R^{\pi^*}\right\|_2, \left\|\mathcal{T}_1 \widehat{h}_R^{\pi^*}\right\|_2^2\right\} + \lambda_n \min\left\{\frac{\eta_n}{2}\left\|\mathcal{T}_2 \widehat{h}_R^{\pi^*}\right\|_2, \left\|\mathcal{T}_2 \widehat{h}_R^{\pi^*}\right\|_2^2\right\} - \lambda_n \mathcal{O}(\eta_n^2). \tag{F.7}$$

Then by following a similar step as in the proof of Theorem 4.15, it follows that

$$\left\|\mathcal{T}_1 \widehat{h}^{\pi^*}\right\|_2 + \left\|\mathcal{T}_2 \widehat{h}^{\pi^*}\right\|_2 = \mathcal{O}(\eta_n).$$

Then combining this result with (F.4), (F.5) and (F.7), we conclude that conditioning on the event $\widetilde{\mathcal{E}}_{R\epsilon}$,

$$\text{Regret}(\widehat{\pi}_R) = \mathcal{O}(\eta_n^{1-\epsilon_R}).$$

Therefore, we complete the proof. □

G Supporting Lemmas for Section 4

In what follows, we present the statement and proofs of the supporting lemmas used in §4.

Lemma G.1 (Directional Derivative)**.** Suppose that Assumption 3.1 and Assumption 4.9 hold. Then for any $h \in \mathcal{H}$,

$$\frac{d\mathcal{T}h^*_\alpha}{dh}[h - h^*_\alpha] = \mathbb{E}[p_{Y|A,X,Z}(h^*_\alpha(A,X))\{h(X,A) - h^*_\alpha(A,X)\} \mid X, Z].$$

Hence $\|h - h^*_\alpha\|_{\text{ps}} = \sqrt{\mathbb{E}[(\mathbb{E}[p_{Y|A,X,Z}(h^*_\alpha(A,X))\{h(X,A) - h^*_\alpha(A,X)\} \mid X, Z])^2]}$.

Proof. Recall that $W(D; h^*_\alpha) = \mathbb{1}\{Y \leqslant h^*_\alpha(A,X)\} - \alpha$. Assumption 4.9 assures that the conditional density $p_{Y \mid A,X,Z}$ exists almost surely. Then

$$\begin{aligned}
\frac{d\mathcal{T}h^*_\alpha}{dh}[h - h^*_\alpha] &= \frac{d\mathbb{E}[W(D; (1-r)h^*_\alpha(A,X) + rh(A,X)) \mid X, Z]}{dr}\Big|_{r=0} \\
&= \frac{d\mathbb{E}[\mathbb{1}\{Y \leqslant h^*_\alpha(A,X) + r[h(A,X) - h^*_\alpha(A,X)]\} - \alpha \mid X, Z]}{dr}\Big|_{r=0} \\
&= \frac{d\mathbb{P}(Y \leqslant h^*_\alpha(A,X)) + r(h(A,X) - h^*_\alpha(A,X) \mid X, Z)}{dr}\Big|_{r=0}.
\end{aligned}$$

By Assumption 4.9, we can write the conditional probability as the integral of the conditional density $p_{Y \mid A,X,Z}$:

$$\begin{aligned}
\frac{d\mathcal{T}h^*_\alpha}{dh}[h - h^*_\alpha] &= \frac{d}{dr}\mathbb{E}[\int_{-\infty}^{h^*_\alpha(A,X)+r(h(A,X)-h^*_\alpha(A,X))} p_{Y \mid A,X,Z}(y)dy \mid X, Z]\Big|_{r=0} \\
&= \mathbb{E}[p_{Y \mid A,X,Z}(h^*_\alpha(A,X) + r[h - h^*_\alpha(A,X)])\{h(X,A) - h^*_\alpha(A,X)\} \mid X, Z]\Big|_{r=0} \\
&= \mathbb{E}[p_{Y \mid A,X,Z}(h^*_\alpha(A,X))\{h(A,X) - h^*_\alpha(A,X)\} \mid X, Z]. \qquad \text{(G.1)}
\end{aligned}$$

Hence $\|h - h^*_\alpha\|_{\text{ps}} = \sqrt{\mathbb{E}[(\mathbb{E}[p_{Y \mid A,X,Z}(h^*_\alpha(A,X))\{h(X,A) - h^*_\alpha(A,X)\} \mid X, Z])^2]}$ by the definition of the pseudometric. □

Proof of Lemma E.1

Proof. Recall that on the event $\mathcal{E}$, $|\mathbb{E}_n[W(D;h)\theta(X,Z)] - \mathbb{E}[W(D;h)\theta(X,Z)]| \leqslant \eta_n(\|\theta\|_2 + \eta_n)$ and $\left|\|\theta\|^2_{n,2} - \|\theta\|^2_2\right| \leqslant \frac{1}{2}\left(\|\theta\|^2_2 + \eta_n^2\right)$. Note we define $\theta^*_h(X,Z)$ as $\arg\min_{\theta\in\Theta}\|\theta - \mathcal{T}h\|_2$. We consider the following two cases:

Case (i) where $\|\theta^*_h\|_2 \geqslant \eta_n$. Let $r = \eta_n/(2\|\theta^*_h\|_2)$. As $r \in [0, 1/2]$ and Θ is star-shaped, $r\theta^*_h \in \Theta$. Thus

$$\mathcal{L}_n(h) \geqslant \mathbb{E}_n[W(D;h)r\theta^*_h(X,Z)] - \frac{1}{2}\|r\theta^*_h\|^2_{n,2}.$$

On the event $\mathcal{E}$, for the second term on the right-hand side,

$$\frac{1}{2}\left\|r\theta_h^*\right\|_{n,2}^2 \leqslant \frac{r^2}{2}\left[\frac{3}{2}\left\|\theta_h^*\right\|_2^2 + \eta_n^2\right] \leqslant \left[\eta_n^2/(4\left\|\theta_h^*\right\|_2^2)\right] \cdot \frac{1}{2}\left[\frac{3}{2}\left\|\theta_h^*\right\|_2^2 + \eta_n^2\right] = \mathcal{O}(\eta_n^2).$$

For the first term regarding the empirical norm of the θ_h^*,

$$\begin{aligned}
\mathbb{E}_n\left[W(D;h)r\theta_h^*(X,Z)\right] &\geqslant \mathbb{E}\left[W(D;h)r\theta_h^*(X,Z)\right] - \eta_n(\left\|r\theta_h^*\right\|_2 + \eta_n) \\
&\geqslant r\mathbb{E}\left[W(D;h)\theta_h^*(X,Z)\right] - \mathcal{O}(\eta_n^2) \\
&= r\mathbb{E}\left[\mathcal{T}h(X,Z)\theta_h^*(X,Z)\right] - \mathcal{O}(\eta_n^2).
\end{aligned}$$

By adding and subtracting $\theta_h^*(X,Z)$, we further have

$$\begin{aligned}
\mathbb{E}_n\left[W(D;h)r\theta_h^*(X,Z)\right] &\geqslant r\mathbb{E}\left[(\mathcal{T}h(X,Z) - \theta_h^*(X,Z) + \theta_h^*(X,Z))\theta_h^*(X,Z)\right] - \mathcal{O}(\eta_n^2) \\
&\geqslant \frac{\eta_n}{2}\left\|\theta_h^*\right\|_2 - \mathcal{O}(\eta_n^2) \geqslant \frac{\eta_n}{2}\left\|\mathcal{T}h\right\|_2 - \mathcal{O}(\eta_n^2),
\end{aligned}$$

where the second inequality follows by Assumption 4.2. Combining both terms, we have

$$\mathcal{L}_n(h) = \sup_{\theta\in\Theta}\mathbb{E}_n\left[W(D;h)\theta(X,Z)\right] - \frac{1}{2}\left\|\theta\right\|_{n,2}^2 \geqslant \frac{\eta_n}{2}\left\|\mathcal{T}h\right\|_2 - \mathcal{O}(\eta_n^2).$$

Case (ii) where $\left\|\theta_h^*\right\|_2 < \eta_n$. We simply choose $r = 1$:

$$\mathcal{L}_n(h) \geqslant \mathbb{E}_n\left[W(D;h)\theta_h^*(X,Z)\right] - \frac{1}{2}\left\|\theta_h^*\right\|_{n,2}^2.$$

We can upper bound the second term by

$$\frac{1}{2}\left\|\theta_h^*\right\|_{n,2}^2 \leqslant \frac{3}{2}\left\|\theta_h^*\right\|_2^2 + \eta_n^2 \leqslant \mathcal{O}(\eta_n^2).$$

Thus the first term can be lower bounded by

$$\begin{aligned}
\mathbb{E}_n\left[W(D;h)\theta_h^*(X,Z)\right] &\geqslant \mathbb{E}\left[W(D;h)\theta_h^*(X,Z)\right] - \eta_n(\left\|\theta_h^*\right\|_2 + \eta_n) \\
&\geqslant \mathbb{E}\left[W(D;h)\theta_h^*(X,Z)\right] - \mathcal{O}(\eta_n^2).
\end{aligned}$$

By adding and subtracting $\theta_h^*(X,Z)$, we have

$$\begin{aligned}
&= \mathbb{E}\left[(\mathcal{T}h(X,Z) - \theta_h^*(X,Z) + \theta_h^*(X,Z))\theta_h^*(X,Z)\right] - \mathcal{O}(\eta_n^2) \\
&\geqslant \left\|\theta_h^*\right\|_2^2 - \mathcal{O}(\eta_n^2) \\
&\geqslant \left\|\mathcal{T}h\right\|_2^2 - \mathcal{O}(\eta_n^2).
\end{aligned}$$

We then complete the proof by combining the above two cases. □

Proof of Lemma E.3

Proof. By definition of $\widehat{h}_R^\pi$, we have

$$v(\widehat{h}_R^\pi, \pi) + \lambda_n \mathcal{E}_n(\widehat{h}_R^\pi) = \inf_{h \in \mathcal{H}} \{v(h, \pi) + \lambda_n \mathcal{E}_n(h)\} \leqslant v(h_\alpha^*, \pi) + \lambda_n \mathcal{E}_n(h_\alpha^*).$$

The inequality holds as $h_\alpha^* \in \mathcal{H}$ from Assumption 4.1. After rearranging the terms and noting that by Theorem 4.5(ii), $\mathcal{E}_n(h_\alpha^*) \leqslant \mathcal{L}_n(h_\alpha^*) \leqslant 13/4 \cdot \eta_n^2$, we have

$$\mathcal{E}_n(\widehat{h}_R^\pi) \leqslant \frac{1}{\lambda_n}\left[v(h_\alpha^*, \pi) - v(\widehat{h}_R^\pi, \pi)\right] + \mathcal{E}_n(h_\alpha^*) \leqslant \frac{2L_h}{\lambda_n} + \frac{13}{4}\eta_n^2 \leqslant 2L_h \eta_n^{1+\epsilon_R} + \frac{13}{4}\eta_n^2.$$

Then by Assumption 4.13, we have $\mathcal{L}_n(\widehat{h}_R^\pi) = \mathcal{E}_n(\widehat{h}_R^\pi) + \inf_{h \in \mathcal{H}} \mathcal{L}_n(h) = \mathcal{O}\left(\eta_n^{1+\epsilon_R}\right)$. □

Proof of Lemma E.2

Proof. Recall the definition of $\mathcal{T}$:

$$\begin{aligned}\mathcal{T}h(X, Z) &:= \mathbb{E}\left[W(D; h) \,|\, X, Z\right] = \mathbb{E}\left[\mathbb{1}\{Y \leqslant h(X, A)\} - \alpha \,|\, X, Z\right] \\ &= \mathbb{E}\left[\int_{-\infty}^{h(A,X)} p_{Y|A,X,Z}(y)dy \,|\, X, Z\right] - \alpha.\end{aligned}$$

Thus, for any $h_1, h_2 \in \mathcal{H}$, by the mean value theorem and Assumption 4.7 we have

$$\begin{aligned}|\mathcal{T}h_1 - \mathcal{T}h_2| &\leqslant \mathbb{E}\left[\sup_{t \in [0,1]} p_{Y|A,X,Z}(h_1(X, A) + t\left[h_2(X, A) - h_1(X, A)\right])\left[h_2(X, A) - h_1(X, A)\right] \,|\, X, Z\right] \\ &\leqslant \mathbb{E}\left[\sup_{t \in [0,1]} p_{Y|A,X,Z}(h_1(X, A) + t\left[h_2(X, A) - h_1(X, A)\right]) \,|\, X, Z\right]\left[\sup_y |h_2(y) - h_1(y)|\right].\end{aligned}$$

Then, by observing that $\sup_{(X,Z) \in \mathcal{X} \times \mathcal{Z}, h \in \mathcal{H}} |\mathcal{T}h| \leqslant 1$, we have

$$\begin{aligned}\|\mathcal{T}h_1\|_2^2 - \|\mathcal{T}h_2\|_2^2 &\leqslant 2\mathbb{E}\left[|\mathcal{T}h_2 - \mathcal{T}h_1|\right] \\ &\leqslant 2\mathbb{E}\left[\sup_{t \in [0,1]} p_{Y|A,X,Z}(h_1(X, A) + t\left[h_2(X, A) - h_1(X, A)\right]) \,|\, X, Z\right]\left[\sup_y |h_2(y) - h_1(y)|\right] \\ &\leqslant 2\mathbb{E}\left[\sup_{t \in [0,1]} p_{Y|A,X,Z}(h_1(X, A) + t\left[h_2(X, A) - h_1(X, A)\right]) \,|\, X, Z\right]\|h_2 - h_1\|_\infty.\end{aligned}$$

Assumption 4.7 now implies the continuity result. □

H Supporting Lemmas for §C.4

In what follows, we present the statement and proofs of the supporting lemmas used in §C.4.

Lemma H.1 (Upper Bound of the Empirical Loss of Negative Controls)**.** If we set $\lambda_n \geqslant \eta_n^{-(1+\epsilon_R)}$ where $0 < \epsilon_R < 1$ is arbitrary, then on the event $\widetilde{\mathcal{E}}$, $\mathcal{E}_n(\widehat{h}_R^\pi) = \mathcal{O}(\eta_n^{1+\epsilon_R})$ and $\mathcal{L}_n(\widehat{h}_R^\pi) = \mathcal{O}(\eta_n^{1+\epsilon_R})$.

Proof. By definition of $\widehat{h}_R^\pi$, we have

$$v(\widehat{h}_{R1}^\pi, \pi) + \lambda_n \mathcal{E}_n(\widehat{h}_R^\pi) = \inf_{h \in \mathcal{H}} \{v(h_1, \pi) + \lambda_n \mathcal{E}_n(h)\} \leqslant v(h_\alpha^*, \pi) + \lambda_n \mathcal{E}_n(h^*).$$

After rearranging the terms and note that by Theorem 4.5(ii), $\mathcal{E}_n(h^*) \leqslant \mathcal{L}_n(h^*) \leqslant (2L_h^2 + 5/4)\eta_n^2$, we have

$$\begin{aligned}\mathcal{E}_n(\widehat{h}_R^\pi) &\leqslant \frac{1}{\lambda_n}\left[v(h_\alpha^*, \pi) - v(\widehat{h}_{R1}^\pi, \pi)\right] + \mathcal{E}_n(h^*) \leqslant \frac{2L_h}{\lambda_n} + \frac{13}{4}\eta_n^2 \\ &\leqslant 2L_h \eta_n^{1+\epsilon_R} + \left(2L_h^2 + \frac{5}{4}\right)\eta_n^2.\end{aligned}$$

Then by Assumption D.10, we have $\mathcal{L}_n(\widehat{h}_R^\pi) = \mathcal{O}\left(\eta_n^{1+\epsilon_R}\right)$. □

Lemma H.2 (Lower Bound of the Empirical Loss for Negative Controls)**.** Suppose that Assumptions D.1 and D.4 hold. On the event $\widetilde{\mathcal{E}}$, for all $h \in \mathcal{H}$, $k = 0, 1$ we have

$$\mathcal{L}_{k,n}(h) \geqslant \min\left\{\frac{\eta_n}{2}\|\mathcal{T}_k h\|_2, \|\mathcal{T}_k h\|_2^2\right\} - \mathcal{O}(\eta_n^2).$$

Proof. This lemma is a direct variation of Lemma E.1. The proof is almost the same. □

Lemma H.3 (Continuity of the Operator for Negative Controls)**.** $\|\mathcal{T}_1 h\|_2^2$ is continuous on $(\mathcal{H}, \|\cdot\|_{\sup})$.

Proof. Recall that $\mathcal{T}_1 h := \mathbb{E}\left[W(D; h_1) - h_2(V, A, X) \,|\, (E, A, X)\right]$. Thus, for any $h_1 = \left(h_1^{(1)}, h_1^{(2)}\right), h_2 = \left(h_2^{(1)}, h_2^{(2)}\right) \in \mathcal{H}$, we have

$$\begin{aligned}|\mathcal{T}_1 h_1 - \mathcal{T}_1 h_2| &= \left|\mathbb{E}\left[W(D; h_1^{(1)}) - h_1^{(2)}(V, A, X) \,|\, (E, A, X)\right] - \mathbb{E}\left[W(D; h_2^{(1)}) - h_2^{(2)}(V, A, X) \,|\, (E, A, X)\right]\right| \\ &\leqslant \left|\mathbb{E}\left[W(D; h_1^{(1)}) - W(D; h_2^{(1)}) \,|\, (E, A, X)\right]\right| \\ &\quad + \left|\mathbb{E}\left[h_1^{(2)}(V, A, X) - h_2^{(2)}(V, A, X) \,|\, (E, A, X)\right]\right|. \qquad \text{(H.1)}\end{aligned}$$

Also, recall that

$$\begin{aligned}\mathbb{E}\left[W(D; h_1) \,|\, E, A, X\right] &= \mathbb{E}\left[\mathbb{1}\{Y \leqslant h_1(X, A)\} - \alpha \,|\, E, A, X\right] \\ &= \int_{-\infty}^{h(A,X)} p(y|E, A, X)dy - \alpha.\end{aligned}$$

Thus, by the mean value theorem, we have

$$\begin{aligned}\left|\mathbb{E}\left[W(D;h_1^{(1)}) - W(D;h_2^{(1)}) \,|\, (E,A,X)\right]\right| &\leqslant \sup_{t\in[0,1]} \left\{ \left| p_{Y\,|\,E,A,X}\big(h_1^{(1)}(X,A) + t[h_2^{(1)}(X,A) - h_1^{(1)}(X,A)]\big)\right| \right. \\ &\qquad \left. \cdot \left| h_2^{(1)}(X,A) - h_1^{(1)}(X,A)\right| \right\} \\ &\leqslant \sup_{t\in[0,1]} \left| p_{Y\,|\,E,A,X}\big(h_1^{(1)}(X,A) + t[h_2^{(1)}(X,A) - h_1^{(1)}(X,A)]\big)\right| \\ &\qquad \cdot \sup_{x,a} \left| h_1^{(1)}(x,a) - h_2^{(1)}(x,a)\right|,\end{aligned}$$

which is upper bounded by

$$\sup_{t\in[0,1]} \left| p_{Y\,|\,E,A,X}\big(h_1^{(1)}(X,A) + t[h_2^{(1)}(X,A) - h_1^{(1)}(X,A)]\big)\right| \cdot \left\| h_1^{(1)} - h_2^{(1)}\right\|_\infty. \tag{H.2}$$

Then, by observing that the $|\mathcal{T}_1 h_1 + \mathcal{T}_1 h_2|$ is upper bounded by $2L_h + 2$, we have

$$\begin{aligned}\|\mathcal{T}h_1\|_2^2 - \|\mathcal{T}h_2\|_2^2 &\leqslant (2L_h+2)\mathbb{E}\left[|\mathcal{T}h_1 - \mathcal{T}h_2|\right] \\ &\leqslant (2L_h+2) \sup_{t\in[0,1]} \left| p_{Y\,|\,E,A,X}\big(h_1^{(1)}(X,A) + t[h_2^{(1)}(X,A) - h_2^{(1)}(X,A)]\big)\right| \cdot \left\| h_1^{(1)} - h_1^{(2)}\right\|_\infty \\ &\quad + (2L_h+2)\left\| h_1^{(2)} - h_2^{(2)}\right\|_\infty,\end{aligned}$$

where the last inequality holds by (H.1) and (H.2). Hence we complete the proof. □

I Concentration Bounds

Throughout this section, we let $\phi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^{H_n}$ and $\psi : \mathcal{X}\times\mathcal{Z} \to \mathbb{R}^{J_n}$ be two feature maps. H_n and J_n represent the dimensions of the embedding spaces, where we assume exponential decay, i.e., H_n and J_n are both $\mathcal{O}(\log n)$. We then define two linear spaces, $\mathcal{H}$ and Θ, as follows:

$$\mathcal{H} = \{\mathcal{X}\times\mathcal{A} \to \beta_1^\top \phi(\cdot) : \beta_1 \in \mathbb{R}^{H_n}, \|\beta_1\|_2 \leqslant C_1, \|\phi\|_{2,\infty} \leqslant 1\}, \tag{I.1}$$

$$\Theta = \{\mathcal{X}\times\mathcal{Z} \to \beta_2^\top \psi(\cdot) : \beta_2 \in \mathbb{R}^{J_n}, \|\beta_2\|_2 \leqslant C_2, \|\psi\|_{2,\infty} \leqslant 1\}. \tag{I.2}$$

The goal of this section is to verify that Condition 4.4 holds with this choice of $\mathcal{H}$ and Θ with $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$. Recall that the event $\mathcal{E}$ in Condition 4.4 is an intersection of two events of the uniform concentration bounds of some function classes. The plan is to first show that the event

$$\left\{ \sup_{w(\cdot)\theta(\cdot)\in\mathcal{Q}} \left|\mathbb{E}_n\left[W(D;h)\theta(X,Z)\right] - \mathbb{E}\left[W(D;h)\theta(X,Z)\right]\right| \leqslant \eta_n(\|\theta\|_2 + \eta_n)\right\}$$

holds with high probability in §I.2. Then we will show that the event

$$\left\{ \left| \|\theta\|_{n,2}^2 - \|\theta\|_2^2 \right| \leqslant \frac{1}{2}\left(\|\theta\|_2^2 + \eta_n^2\right), \forall \theta \in \Theta \right\}$$

holds with high probability in §I.3. Finally, we take a union bound of the two events to complete the proof of Condition 4.4.

I.1 Concepts from the Empirical Process Theory

To characterize the concentration bound, we first need to introduce some concepts in the empirical process theory.

Bracketing number. Given two functions f_1 and f_2 such that $\|f_1 - f_2\| \leqslant t$, A t-bracket $[f_1, f_2]$ is a subset of functions of the real-valued function class $\mathcal{F}$ on $\mathcal{X}$ satisfying for all $f \in [f_1, f_2], \forall x \in \mathcal{X}, f_1(x) \leqslant f(x) \leqslant f_2(x)$. The bracketing number $N_{[]}(t; \mathcal{F}, \|\cdot\|_2)$ of a function class $\mathcal{F}$ with respect to the norm $\|\cdot\|_2$ is the smallest number of t-brackets needed to cover $\mathcal{F}$.

Covering number. The covering number $N(t; \mathcal{F}, \|\cdot\| \|_\bullet)$ of a function class $\mathcal{F}$ with respect to some norm $\|\cdot\| \|_\bullet$ is the smallest number of t-balls in $\mathcal{F}$ needed to cover $\mathcal{F}$.

Localized population Rademacher complexity. The localized population Rademacher complexity with respect to $\{X_i\}_{i=1}^n$ and function class $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\eta; \mathcal{F}) = \mathbb{E}_{\{\varepsilon_i\}_{i=1}^n, \{X_i\}_{i=1}^n} \left[\sup_{f \in \mathcal{F}, \|f\|_2 \leqslant \eta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right].$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables.

Critical radius. Suppose $\mathcal{F}$ is a real-valued function class on $\mathcal{X}$ such that $\forall f \in \mathcal{F}, \|f\|_\infty \leqslant c$ for some constant $c \geqslant 0$. The critical radius of $\mathcal{F}$ is the largest possible η such that $\mathcal{R}_n(\eta; \mathcal{F}) \leqslant \eta^2/c$.

I.2 Concentration bound of the Function Class $\mathcal{Q}$ Using Bracketing Number

Recall that we define the function class $\mathcal{Q}$ as $\mathcal{Q} = \{W(\cdot; h) \times \theta(\cdot) : h \in \mathcal{H}, \theta \in \Theta\}$. In this subsection, we show that the first event of Condition 4.4 holds with high probability.

Theorem I.1 (Rate of Tail Bound of the Function Class $\mathcal{Q}$.)**.** If we choose $\mathcal{H}$ as (I.1) and Θ as (I.1), then the event $\{|\mathbb{E}_n [W(D; h)\theta(X, Z)] - \mathbb{E} [W(D; h)\theta(X, Z)]| \leqslant \eta_n (\|\theta\|_2 + \eta_n)\}$ holds with probability $1 - \xi$ for some $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$.

Proof. We proceed with the help of the following two lemmas.

Lemma I.2 (Bracketing Number of the Function Class $\mathcal{Q}$)**.** For every t small enough, under Assumption 4.7, there exists $C_3 > 0$ such that

$$\log N_{[]}(t; \mathcal{Q}, \|\cdot\|_2) \leqslant A_n \cdot \log(C_3/t), \tag{I.3}$$

where $A_n = 4(H_n + J_n)$.

Proof. See §J.1 for a detailed proof. □

Lemma I.3 (Lemma 3.11. in Hu et al. (2020))**.** Let $\widetilde{\mathcal{Q}}$ be a class of functions uniformly bounded by one in $\|\cdot\|_2$ and q_0 be a fixed element in $\widetilde{\mathcal{Q}}$. Let $\widetilde{Q}(t) = \{q \in \widetilde{\mathcal{Q}} : \|q - q_0\|_2) \leqslant t\}$. Suppose $\log N_{[]}(t; \widetilde{\mathcal{Q}}, \|\cdot\|_2) \leqslant A_n \log(C/t)$ for some $C > 0$. Then there exist universal positive constants C_4, C_5, C_6, and C_7 such that for any $\xi > 0$,

$$\mathbb{P}\left(\sup_{\substack{q\in\widetilde{\mathcal{Q}};\\ \|q-q_0\|_2>\sqrt{\frac{A_n}{n}}}} \frac{|(\mathbb{E}_n - \mathbb{E})(q)|}{\sqrt{\frac{A_n}{n}}\|q-q_0\|_2 \log(C\|q-q_0\|_2^{-1})} \leqslant \frac{C_4}{A_n}\log(C_5/\xi) \right) \geqslant 1 - \xi/2. \tag{I.4}$$

Moreover, for $q \in \widetilde{\mathcal{Q}}(\sqrt{\frac{A_n}{n}})$, we have

$$\mathbb{P}\left(\sup_{q\in\widetilde{\mathcal{Q}}(\sqrt{\frac{A_n}{n}})} |(\mathbb{E}_n - \mathbb{E})(q)| \leqslant \frac{C_6\log(C_7/\xi)}{A_n \log^2(Cn/A_n)} \frac{A_n}{n}\log(Cn/A_n) \right) \geqslant 1 - \xi/2. \tag{I.5}$$

This lemma is a simple variant of Lemma 5.13 in van de Geer (2009).

By Lemma I.2, we can substitute q_0 as the zero function and $\widetilde{\mathcal{Q}} = \mathcal{Q}/C_1$ in Lemma I.3. Then the event in (I.4) concerns the supremum over $q \in \mathcal{Q}$ and $\|q\|_2 \geqslant C_1\sqrt{A_n/n}$. Since when $\|q\|_2 \geqslant C_1\sqrt{A_n/n}$, $\log(C_3C_1/\|q\|_2) \leqslant \log(C_3\sqrt{n/A_n})$, (I.4) becomes:

$$\mathbb{P}\left(\sup_{\substack{q\in\mathcal{Q};\\ \|q\|_2>C_1\sqrt{\frac{A_n}{n}}}} \frac{|(\mathbb{E}_n - \mathbb{E})(q)|}{\sqrt{\frac{A_n}{n}}\|q\|_2 \log\left(C_3\sqrt{\frac{n}{A_n}}\right)} \leqslant \frac{C_4}{A_n}\log(C_5/\xi) \right) \geqslant 1 - \xi/2. \tag{I.6}$$

Moreover, the restriction $q \in \widetilde{\mathcal{Q}}(\sqrt{A_n/n})$ of the event in (I.5) is replaced with $q \in \mathcal{Q}(C_1\sqrt{A_n/n})$. We now set

$$\eta_n = \max\left\{ \sqrt{\frac{A_n}{n}}\log\left(C_3\sqrt{\frac{n}{A_n}}\right)\cdot\frac{C_4}{A_n}\cdot\log\left(\frac{C_5}{\xi}\right),\ \left[\frac{C_1A_n}{n}\cdot\frac{C_6\log(C_7/\xi)}{A_n\log^2(C_3n/C_1A_n)}\cdot\log(C_3n/C_1A_n)\right]^{1/2} \right\}.$$

Then combine (I.5) and (I.6) and take the union bond, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w(\cdot),\theta(\cdot)\in\mathcal{Q}} \left|\mathbb{E}_n\left[W(D;h)\theta(X,Z)\right]-\mathbb{E}\left[W(D;h)\theta(X,Z)\right]\right| \leqslant \eta_n\left(\|\theta\|_2+\eta_n\right)\right)\\
&\geqslant \mathbb{P}\left(\sup_{w(\cdot),\theta(\cdot)\in\mathcal{Q}\left(C_1\sqrt{\frac{A_n}{n}}\right)} \left|\mathbb{E}_n\left[W(D;h)\theta(X,Z)\right]-\mathbb{E}\left[W(D;h)\theta(X,Z)\right]\right| \leqslant \eta_n^2,\right.\\
&\qquad\left.\sup_{\|w(\cdot)\theta(\cdot)\|_2\geqslant C_1\sqrt{\frac{A_n}{n}}} \left|\mathbb{E}_n\left[W(D;h)\theta(X,Z)\right]-\mathbb{E}\left[W(D;h)\theta(X,Z)\right]\right| \leqslant \eta_n\|\theta\|_2\right)\\
&\geqslant 1-\xi.
\end{aligned}$$

We complete the proof by noticing that $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$ by choice. □

I.3 Concentration Bound Using Critical Radius

We now demonstrate that there exists $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$ such that the event $\{|\,\|\theta\|_{n,2}^2 - \|\theta\|_2^2\,| \leqslant \frac{1}{2}\left(\|\theta\|_2^2+\eta_n^2\right), \forall\theta\in\Theta\}$ holds with high probability when Θ is chosen to be the linear space I.2. Subsequently, we can then take the union bound to verify Condition 4.4.

Lemma I.4. For any $\xi > 0$, there exists $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$ such that the event $\{|\,\|\theta\|_{n,2}^2 - \|\theta\|_2^2\,| \leqslant \frac{1}{2}\left(\|\theta\|_2^2+\eta_n^2\right), \forall\theta\in\Theta\}$ holds with probability $1-\xi$ when Θ is chosen to be the linear space (I.2).

Proof. We proceed with the help of two lemmas:

Lemma I.5 (Lemma F.4 in Chen et al. (2023))**.** The covering numbers $N(t;\mathcal{H},\|\cdot\|_\infty)$ and $N(t;\Theta,\|\cdot\|_\infty)$ are upper bounded by $(1+2C_1/t)^{H_n}$ and $(1+2C_2/t)^{J_n}$, respectively.

Lemma I.6 (Theorem 14.1 in Wainwright (2019))**.** Let t_n be the critical radius of Θ. There exist universal positive constants k_0 and k_1 such that for any $\eta_n \geqslant t_n + k_0\sqrt{\log(k_1/\xi)/n}$, we have

$$|\,\|\theta\|_n^2 - \|\theta\|_2^2\,| \leqslant \frac{1}{2}\|\theta\|_2^2 + \frac{1}{2}\eta_n^2, \quad \forall\theta\in\Theta$$

with probability at least $1-\xi$.

By Lemma I.5, we can substitute Θ for $\mathcal{F}$ in J.3. Hence the critical radius of Θ is bounded by $\mathcal{O}(\sqrt{(J_n\log n)/n}) = \widetilde{\mathcal{O}}(n^{-1/2})$. We then use Lemma I.6 to complete the proof. □

Theorem I.7 (Main Theorem of Concentration Bound)**.** Condition 4.4 holds when choosing $\mathcal{H}$ and Θ as in (I.1) and (I.2) with some $\eta_n = \widetilde{\mathcal{O}}(n^{-1/2})$.

Proof. By Theorem I.1, we have

$$\mathbb{P}\left(\left|\mathbb{E}_n\left[W(D;h)\theta(X,Z)\right]-\mathbb{E}\left[W(D;h)\theta(X,Z)\right]\right|\leqslant\eta_n\left(\|\theta\|_2+\eta_n,\forall h\in\mathcal{H},\forall\theta\in\Theta\right)\right)\geqslant 1-\xi.$$

Combine with Lemma I.4 and take the union bound, and we complete the proof. □

J Supporting Lemmas of §I

In what follows, we introduce and prove supporting lemmas concerning the bracketing and covering numbers of certain function classes of interest, as discussed in §I. These lemmas play a crucial role in verifying Condition 4.4.

J.1 Proof of Lemma I.2

Proof. We prove this Lemma with the help of the two Lemmas below.

Lemma J.1. The bracketing numbers $N_{[]}(t;\Theta,\|\cdot\|_2)$ is upper bounded by $(1+4C_1/t)^{J_n}$.

Proof. Note that $\mathcal{H}$ is indexed by the set $\widetilde{\beta}=\{\beta_2\in\mathbb{R}^{J_n}:\|\beta_2\|_2\leqslant C_2\}$. Let $F(x,z)\equiv 1$. For any $\beta_2^T\psi(\cdot),\widetilde{\beta}_2^T\psi(\cdot)\in\Theta$, i.e., $\beta_2,\widetilde{\beta}_2\in\widetilde{\beta}$, we have

$$|\beta_2^T\psi(x,z)-\widetilde{\beta}_2^T\psi(x,z)|\leqslant\left\|\beta_2-\widetilde{\beta}_2\right\|_2\|\psi\|_{2,\infty}F(x,z)\leqslant\left\|\beta_2-\widetilde{\beta}_2\right\|_2$$

almost surely in (X,Z). So by Theorem 2.7.11 in van der vaart and Wellner (2013),

$$N_{[]}(t;\Theta,\|\cdot\|_2)\leqslant N(t/2;\widetilde{\beta},\|\cdot\|_2)\leqslant N(t/2;\Theta,\|\cdot\|_2)\leqslant N(t/2;\Theta,\|\cdot\|_\infty)\leqslant(1+4C_2/t)^{J_n},$$

where the last step leverages Lemma I.5. □

Lemma J.2 (Theorem 3 & Example 5.1 in Chen et al. (2003))**.** Let the indicator function class be

$$\mathcal{W}:=\{W(\cdot;h):h\in\mathcal{H}\}.$$

Suppose Assumption 4.7 holds. Then the bracketing number $N_{[]}(t;\mathcal{W},\|\cdot\|_2)$ is upper bounded by $\left(1+8C_1K^2/t^2\right)^{H_n}$ for t small enough, where $K>0$ uniformly upper bounds $\sup_y p_{Y\,|\,A,X,Z}(y)$ for almost all (A,X,Z).

Proof. In the proof of Theorem 3 in Chen et al. (2003), for t small enough, with Assumption 4.7:

$$N_{[]}(t;\mathcal{F}_2,\|\cdot\|_2)\leqslant N((t/2K)^{\frac{1}{s}};\Theta',\|\cdot\|_\infty)N((t/2K)^{\frac{1}{s}};\mathcal{H},\|\cdot\|_\infty).$$

In our case of indicator functions, as also indicated in Example 5.1 in Chen et al. (2003), choose $s=1/2$ and $\mathcal{F}_2=\mathcal{W}$. Let $\Theta'=\{\alpha\}$ and combine with Lemma I.5, we have

$$N_{[]}(t,\mathcal{W},\|\cdot\|_2)\leqslant 1\cdot N((t/2K)^2;\mathcal{H},\|\cdot\|_\infty)\leqslant\left(1+\frac{8C_1K^2}{t^2}\right)^{H_n}.$$

Therefore, we complete the proof. □

We prove the upper bound of the function class $\mathcal{Q}$ by combining the upper bounds of the function classes Θ and $\mathcal{W}$. Consider $t \in (0,1)$ small enough, we have

$$\begin{aligned}
\log N_{[]}(t, \mathcal{Q}, \|\cdot\|_2) &\leqslant \log N_{[]}(t/2; \mathcal{W}, \|\cdot\|_2) + N_{[]}(t/2; \Theta, \|\cdot\|_2) \\
&\leqslant H_n \log(1 + 8C_1K^2/t^2) + J_n \log(1 + 4C_2/t) \\
&\leqslant (H_n + J_n) \log\left((1 + 8C_1K^2/t^2)(1 + 4C_2/t^2)\right) \\
&\leqslant 2(H_n + J_n) \log\left(\frac{1 + \max\{8C_1K^2, 4C_2\}}{t^2}\right),
\end{aligned}$$

where the second inequality holds by Lemma J.1 and Lemma J.2. We then let $A_n = 4(H_n + J_n)$ and $C_3 = \sqrt{1 + \max\{8C_1K^2, 4C_2\}}$ to complete the proof. □

Lemma J.3 (Lemma F.7 in Chen et al. (2023))**.** If the covering number of C_2-uniformly bounded function class $\mathcal{F}$ satisfies $\log N(t; \mathcal{F}, \|\cdot\|_\infty) \leqslant A \log(1 + 2C_2/t)$ for some constant A, the critical radius of $\mathcal{F}$ is upper bounded by $\mathcal{O}(\sqrt{(A \log n)/n})$.