Journal of Machine Learning Research 25 (2024) 1-56 Submitted 8/22; Revised 9/24; Published 10/24

Instrumental Variable Value Iteration for Causal Offline Reinforcement Learning

Luofeng Liao* LL3530@COLUMBIA.EDU
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

Zuyue Fu* ZUYUEFU2022@U.NORTHWESTERN.EDU
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

Zhuoran Yang ZHUORAN.YANG@YALE.EDU
Department of Statistics and Data Science
Yale University
New Haven, CT 06520, USA

Yixin Wang YIXINW@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Dingli Ma DINGLI98@UW.EDU
Department of Information Systems and Operations Management, Michael G. Foster School of Business
University of Washington
Seattle, WA 98195, USA

Mladen Kolar MKOLAR@MARSHALL.USC.EDU
Department of Data Sciences and Operations, Marshall School of Business
University of Southern California
Los Angeles, CA 90089, USA

Zhaoran Wang ZHAORANWANG@GMAIL.COM
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

Editor: Eric Laber

Abstract

In offline reinforcement learning (RL) an optimal policy is learned solely from a priori collected observational data. However, in observational data, actions are often confounded by unobserved variables. Instrumental variables (IVs), in the context of RL, are the variables whose influence on the state variables is all mediated by the action. When a valid instrument is present, we can recover the confounded transition dynamics through observational data. We study a confounded Markov decision process where the transition dynamics admit an additive nonlinear functional form. Using IVs, we derive a conditional moment restriction through which we can identify transition dynamics based on observational data. We propose a provably efficient IV-aided Value Iteration (IVVI) algorithm based on a primal-dual reformulation of the conditional moment restriction. To our knowledge, this is the first provably efficient algorithm for instrument-aided offline RL.

Keywords: instrumental variables, reinforcement learning, causal inference

1. Introduction

In reinforcement learning (RL) (Sutton and Barto, 2018), an agent maximizes its expected total reward by sequentially interacting with the environment. RL algorithms have been applied in the healthcare domain to dynamically suggest optimal treatments for patients with certain diseases (Raghu et al., 2017; Komorowski et al., 2018; Futoma et al., 2018; Namkoong et al., 2020; Guez et al., 2008; Parbhoo et al., 2017; Prasad et al., 2017). One of the main concerns of working with observational data, especially for RL applications in healthcare, is the confounding caused by unobserved variables. Because available data may not contain measurements of important prognostic variables that guide treatment decisions or heuristic information such as visual inspection or discussions with patients during each treatment period, there exist variables that affect both treatment decisions and the next stage health status of patients. See Brookhart et al. (2010) for a detailed discussion of sources of confounding in healthcare datasets.

Another motivating example for this work is recommender systems. In movie recommendation systems, the platform collects users' viewing history and movie ratings. It is desirable to learn from the collected datasets a movie recommendation policy that fits users' preferences and results in high movie ratings. However, there are often factors that affect users' action (watch movies or not) and movie preference. For example, the director or the star of the movie (Wang et al., 2020).

Instrumental variables (IVs) are a well-known tool in econometrics and causal inference to identify causal effects in the presence of unobserved confounders (UCs). Informally, a variable Z is an IV for the causal effect of the treatment variable X on the outcome variable Y, if (i) it is correlated with X, and (ii) Z affects only Y through X, and (iii) Z should be exogenous, e.g., Z is independent of unobserved confounders. We provide several concrete use cases below, beginning with a recommendation system application.

Example 1 ((Recommendation as an IV, MovieLens 1M data)). *In recommender systems, we could model users' experience as the outcome variable, and watching some movie as the treatment. The goal is to identify a sequence of movies that improve user experience when the user actually watches these movies. Conditional on a user, when the recommendation is sufficiently randomized, the recommendation itself can be used as an IV to deconfound the effect of a movie to user experience. We discuss this application with a semi-synthetic dataset based on the MovieLens 1M dataset (Harper and Konstan, 2015) in Section 5.3.*

IVs are also commonly used in the healthcare domain to identify the effects of a treatment or intervention on health outcomes. There are some common sources of IVs in the medical literature, such as preference-based IVs (see Example 3), distance to a specialty care provider (see Example 2), and genetic variants (Baiocchi et al., 2014). Such a wide range of potential use of IVs in these healthcare sequential decision-making settings is a key motivator of the paper.

Example 2 ((Differential travel time as an IV, NICU data)). *Lorch et al. (2012), Michael et al. (2023), and Chen and Zhang (2023) studied the effect of delivery on neonatal mortality in high-level neonatal intensive care units (NICU), using the same differential travel time as an IV. The goal is to design a neonatal regionalization system that designates hospitals according to the level of care that infants need. The available dataset has ∼180,000 records of mothers who delivered exactly two births during 1995 and 2009 in Pennsylvania and relocated at the second delivery. In Figure 1 we present a possible causal DAG for the NICU application. UCs are present due to mothers' self-selection effects or unrecorded side information on which the physicians base the NICU suggestion. The differential travel time to the closest high-level NICU versus low-level NICU serves as a valid IV, since it affects the choice of the mother's NICU and does not impact clinical outcomes through other means. A neonatal regionalization system (Figure 1, bottom panel) designates the NICU solely based*

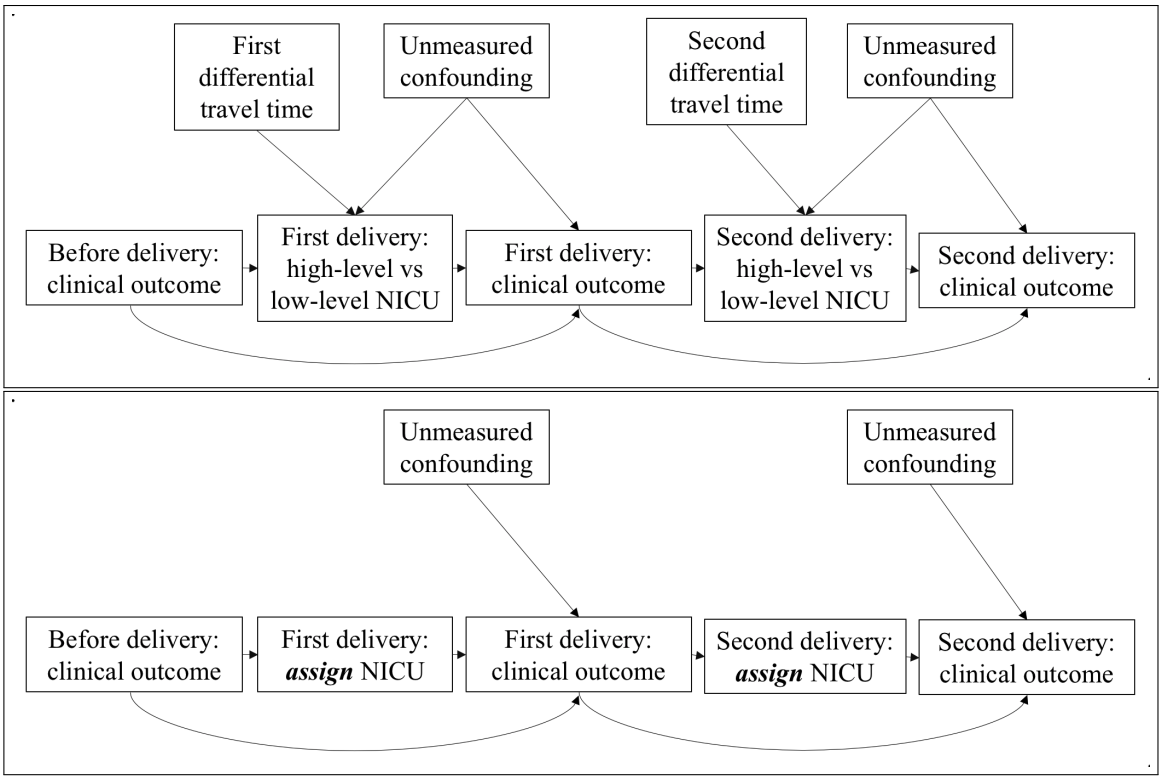


Figure 1: The NICU application, adapted from Chen and Zhang (2023, Figure 1). Sufficient covariates have been conditioned on. Top panel: DAG representing data generation process where UCs are present. Bottom panel: DAG representing a prenatal regionalization system in action.

on the clinical outcome at the previous stage (since differential travel time does not affect the clinical outcome anymore once we actually assign NICU, and confounders remain unobserved), removing arrows that point to the decision of the NICU in the DAG presented in the upper panel.

Example 3 ((Preference-based IV, MIMIC-III data))**.** *For example, the work of Brookhart and Schneeweiss (2007) discusses the use of preference-based IVs. They assume that different healthcare providers, at the level of geographic regions, hospitals, or individual physicians, have different preferences on how medical procedures are performed. Then preference-based IVs are variables that represent the variation in these healthcare providers. In the context of sepsis management by applying RL (Komorowski et al., 2018) on the MIMIC-III dataset (Johnson et al., 2016), the effect of doses of intravenous fluids and vasopressors (X) on the health status of patients (Y) is likely to be confounded by unrecorded severity level of comorbidities. Then a physician's preference for prescribing vasopressors (Z) is a potentially valid IV since it affects directly the actual doses given (X), but is unlikely to affect the next-stage health status through other causes of Y.*

We summarize three aspects of offline sequential datasets often encountered by RL practitioners: (i) there is a large amount of logged data where the actual effects of action on the outcome are confounded, (ii) a valid IV is, in some situations, available, and (iii) it is expensive or unethical to do experimentation and then inspect the actual performance of a target policy. We ask

When a valid IV is present, can we design a provably efficient offline RL algorithm using only confounded observational data?

We answer this question affirmatively. We formulate the sequential decision-making process in the presence of both IVs and UCs through a model that we call Confounded Markov Decision Process with Instrumental Variables (CMDP-IV). We then propose an IV-aided Value Iteration (IVVI) algorithm to recover the optimal policy through a model-based approach. Our contribution is threefold. First, under the additive UC assumption, we derive a conditional moment restriction through which we point identify transition dynamics. Second, we reformulate the conditional moment restriction as a primal-dual optimization problem and propose an estimation procedure that enjoys computational and statistical efficiency jointly. Finally, we show that the sample complexity of recovering an ϵ-optimal policy using observational data with IVs is $O(\mu_{\text{IV}}^{-4}\mu_B^{-2.5}H^4 d_x \epsilon^{-2})$, where $0 < \mu_{\text{IV}} < 1$ quantifies the strength of the IV, μ_B is the minimum eigenvalue of the dual feature covariance matrix, quantifying the compatibility of the dual linear function space and the IV, H is the horizon of the MDP, and d_x is the dimension of states. To the best of our knowledge, this is the first result on sample complexity for an IV-aided offline RL.

Several results developed in the paper are worth noting. We propose a stochastic approximation estimator for nonparametric IV problem, which is jointly computationally and statistically efficient. We are also among the first to study offline RL in multi-stage settings with continuous actions in the face of unobserved confounding and continuous IVs. Our results on solving a stochastic quadratic saddle-point problem may be of independent interest.

1.1 Related Work

Identification of Causal Estimand in Sequential Settings RL in the presence of UCs has attracted increasing attention. One major difficulty of working with unobserved confounders is the issue of identification. When unobserved confounders are present, causal effects of actions are not identifiable from data without further assumptions. In these settings, several approaches are available. The first one is the sensitivity-analysis based approach (Rosenbaum, 2002), where we posit additional sensitivity assumptions on how strong the unobserved confounding can possibly be. These sensitivity assumptions enable partial identification of the causal quantity. This approach is employed by a series of work by Kallus et al. (2019); Kallus and Zhou (2020, 2021); Namkoong et al. (2020). The second approach is to assume access to other auxiliary variables that can enable point or partial identification. We adopt the second approach in this work, by assuming the access to instrumental variables. Under an additive UC assumption (see Eq. 2.6), instrumental variables can enable **point identification** of the structural quantity through conditional moment restriction (along with certain completeness assumptions), allowing us to work with continuous actions and continuous IVs. For example, in the NICU application, differential travel time (the IV) is a continuous quantity. Note that several other related works also study the use of instrumental variables (Pu and Zhang, 2021; Chen and Zhang, 2023). These works, and in particular the work by Chen and Zhang (2023), rely on partial identification bounds in the fully nonparametric IV setting (Manski, 1990; Balke and Pearl, 1994). These bounds are only available for binary IVs or binary treatments, restricting the use of their algorithms in many real-world scenarios where the IV is continuous. A continuous IV like the differential travel time must be dichotomized if one were to apply these algorithms.

Dynamic treatment regime (DTR) DTRs (Murphy, 2003; Chakraborty and Moodie, 2013; Chakraborty and Murphy, 2014) are a popular model for sequential decision making. DTR learning differs from RL in that it does not require the Markov assumption and the quantity of interests is an optimal adaptive dynamic policy that makes its decision based on all information available prior to the decision point. However, unobserved confounding is often expected in observational data, and yet few works handle UCs in DTR learning. A concurrent work by Chen and Zhang (2023)

study the policy improvement problem in the presence of UCs, using partial identification results of causal quantities with IVs (Manski, 1990; Balke and Pearl, 1994). However, these identification results often apply to binary treatments or binary IVs, restricting their use in many real-world scenarios where the IV is continuous. In our work, the transition function is point-identified under the additive UC assumption. This enables us to work with continuous actions and continuous IVs.

RL in the presence of UCs. Zhang and Bareinboim (2016) formulate the MDP with unobserved confounding using the language of structural causal models. Lu et al. (2018) study a model-based RL algorithm in a combined online and observation setting. They propose a structural causal model for the confounded MDP and estimate the structural function with neural nets using the observational data. Buesing et al. (2018) propose a model-based RL algorithm in the evaluation setting that learns the optimal policy for a partially observable Markov decision process (POMDP). Oberst and Sontag (2019) propose a class of structural causal models (SCMs) for the data generating process of POMDPs and then discuss identification of counterfactuals of trajectories in the SCMs. Tennenholtz et al. (2020) study offline policy evaluation in POMDP. Their identification strategy relies on the identification results of proxy variables in causal inference (Miao et al., 2018). Zhang and Bareinboim (2019, 2020) study the dynamic treatment regime and propose an algorithm to recover optimal policy in the online RL setting that is based on partial identification bounds of the transition dynamics, which they use to design an online RL algorithm. Namkoong et al. (2020) study offline policy evaluation when UCs affect only one of the many decisions made. They work with a partially identified model and construct partial identification bounds of the target policy value. Bennett et al. (2021) study off-policy evaluation in infinite horizon. Their method relies on estimation of the density ratio of the behavior policy and target policy through a conditional moment restriction. Kallus and Zhou (2020) study off-policy evaluation in infinite horizon. They characterize the partially identified set of policy values and compute bounds on such a set. Kallus and Zhou (2018, 2021) study policy improvement using sensitivity analysis.

Primal-dual estimation of nonparametric IV (NPIV) Typical nonparametric approaches to IV regression include smoothing kernel estimators and sieve estimators Newey and Powell (2003); Carrasco et al. (2007); Chen and Christensen (2018); Darolles et al. (2011), and very recently, reproducing kernel Hilbert space-based estimators Singh et al. (2019); Muandet et al. (2020). However, traditional nonparametric methods are not scalable and thus not suitable for modern-day RL datasets.

Our proposed method builds on a recent line of work that investigates primal-dual estimation of NPIV (Dai et al., 2017; Lewis and Syrgkanis, 2018; Bennett et al., 2019; Muandet et al., 2020; Dikkala et al., 2020; Liao et al., 2020).

This paper differs from previous works in primal-dual estimation of NPIV in two aspects. First, we solve the NPIV problem through a stochastic approximation (SA) approach (Robbins and Monro, 1951). The SA approach is an online procedure in the sense it updates the estimate upon receiving new data points. This is a more desirable framework for practical RL applications. For example, in business application of RL, data is logged following business as usual, streaming into the database system. New technology such as wearable devices allows real-time collection of health information, medical decisions and their associated outcomes. Faced with large amounts of data, practitioners typically prefer algorithms that process new data points in real time; see Remark 13 for a detailed comparison with the sample average approximation approach. Our stochastic approximation approach to NPIV problem tackles computational error and statistical error jointly and is well-suited for streaming data.

Second, despite that the stochastic saddle-point problem is not strongly-convex-strongly-concave, we show a fast rate of $O(1/T)$ can be attained by a simple stochastic gradient descent-ascent algorithm.

1.2 Notation

We use $\|\cdot\|_2$ to denote the ℓ_2-norm of a vector or the spectral norm of a matrix, and use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. For vectors a, b of the same length, let $a \cdot b$ denote the inner product. We denote by $\Delta(\mathcal{M};\mathcal{N})$ the set of distributions on $\mathcal{M}$ indexed by elements in $\mathcal{N}$. For a real symmetric matrix A, let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ be its largest and smallest eigenvalues, respectively. For any positive integer n, we define $[n] = \{1, \ldots, n\}$. For any bounded function $\varphi\colon \mathcal{X} \to \mathbb{R}^{d_\varphi}$, we define the linear function space spanned by φ as $\mathcal{H}_\varphi = \{\theta \cdot \varphi\colon \theta \in \mathbb{R}^{d_\varphi}\}$. For any function $f = \theta \cdot \varphi \in \mathcal{H}_\varphi$, we denote by $\|f\|_\varphi = \|\theta\|_2$ its norm.

2. Problem Setup

We formulate the problem in this section. We first define instrumental variables (IVs) in Section 2.1 as a preliminary. In Section 2.2.1, we describe the *evaluation setting*, where we test the performance of our learned policy. In Section 2.2.2, we describe the *observation setting* in which we collect the observational data to learn a policy. Our goal is then to recover the optimal policy for the evaluation setting, using only data collected in the observation setting.

2.1 Preliminaries: Instrumental Variables

We define confounders and IVs as follows.

Definition 1 (Confounders and Instrumental Variables, Pearl 2009)**.** *A variable ε is a confounder relative to the pair (X, Y) if (X, Y) are both caused by ε. A variable Z is an IV relative to the pair (X, Y), if it satisfies the following two conditions: (i) Z is independent of all variables that have influence on Y and are not mediated by X; (ii) Z is not independent of X.*

Figure 2 (left panel) illustrates a typical causal directed acyclic graph (DAG) for an IV, where Z is the IV relative to the pair (X, Y), and ε is the UC relative to the pair (X, Y). The the DAG in Figure 2 (left) can also be characterized by $X = g(Z, \varepsilon)$ and $Y = f(X, \varepsilon)$ given independent Z and ε, where f and g are two deterministic functions.

2.2 CMDP-IV

We first introduce a type of finite-horizon Markov Decision Process (MDP) in the observation setting with UCs and IVs, which we term *Confounded Markov Decision Process with Instrumental Variables* (CMDP-IV). CMDP-IV is a natural extension of the IV model introduced in Section 2.1 to the multi-stage decision making process.

A CMDP-IV is defined as a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{U}, H, r; \xi_0, \mathcal{P}_e, \mathcal{P}_z, F^*, \pi_b)$, where the sets $\mathcal{S} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{A}$ are state and action spaces; the set $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ is the space of IVs; the set $\mathcal{U} \subseteq \mathbb{R}^{d_x}$ is the space of UCs; the integer H is the length of each episode; and $r = \{r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]\}_{h=1}^H$ is the set of deterministic reward functions, where r_h is the reward function at the h-th step. For simplicity of presentation, we assume that the reward function r_h is known for any $h \in [H]$. Furthermore, $\xi_0 \in \Delta(\mathcal{S})$ is the initial state distribution, $\mathcal{P}_e = \mathcal{N}(0, \sigma^2 I_{d_x})$ is the distribution of UCs, and $\mathcal{P}_z$ is the distribution of IVs. The function $F^* : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is a deterministic transition function and $\pi_b = \{\pi_{b,h} \in \Delta(\mathcal{A}; \mathcal{S}, \mathcal{Z}, \mathcal{U})\}_{h=1}^H$ is the behavior policy, where $\pi_{b,h}$ is the behavior policy at the h-th step.

2.2.1 Evaluation setting: Bellman Equations and Performance Metric

We now introduce the evaluation setting of CMDP-IV. The evaluation setting is the same as the usual RL setup (Sutton and Barto, 2018): we want to find an optimal policy in the MDP.

For a policy $\pi = \{\pi_h \in \Delta(\mathcal{A}; \mathcal{S})\}_{h=1}^H$, given an initial state $x_1 \sim \xi_0$, for any $h \in [H]$, the dynamics in an evaluation setting at the h-th step is

$$a_h \sim \pi_h(\cdot \,|\, x_h) \;, \quad x_{h+1} = F^*(x_h, a_h) + e_h \;, \tag{2.1}$$

where $\{e_h\}_{h=1}^H \overset{\text{iid}}{\sim} \mathcal{P}_e$ is the sequence of Gaussian innovations. The episode terminates if we reach the state x_{H+1}. For simplicity, for any $F : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_x}$ we define the following transition kernel

$$\mathcal{P}_F(\cdot \,|\, x_h, a_h) = \mathcal{N}(F(x_h, a_h), \sigma^2 I_{d_x}) \;. \tag{2.2}$$

We define the value function and the Q-function of a policy under the evaluation setting Eq. (2.1). For any $h \in [H]$, given any policy π_h at the h-th step, we define its value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ and its Q-function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows,

$$V_h^\pi(x) := \mathbb{E}_\pi\Big[\sum_{i=h}^H r_i(x_i, a_i) \,\Big|\, x_h = x\Big] \;, \;\; Q_h^\pi(x, a) := \mathbb{E}_\pi\Big[\sum_{i=h}^H r_i\left(x_i, a_i\right) \Big| x_h = x, a_h = a\Big] \;. \tag{2.3}$$

Here, the expectation $\mathbb{E}_\pi$ is taken with respect to the randomness of the state-action sequence $\{(x_i, a_i)\}_{i=h}^H$, where the action a_i follows the policy $\pi_i(\cdot \,|\, x_i)$ and the next state x_{i+1} follows the transition kernel $\mathcal{P}_{F^*}(\cdot \,|\, x_i, a_i)$ defined in Eq. (2.2) for any $i \in \{h, h+1, \ldots, H\}$.

An optimal policy π^* gives the optimal value $V_h^*(x) = \sup_\pi V_h^\pi(x)$ for any $(x, h) \in \mathcal{S} \times [H]$. We assume that such an optimal policy π^* exists. For a given policy $\pi = \{\pi_h \in \Delta(\mathcal{A}; \mathcal{S})\}_{h=1}^H$, its suboptimality compared to the optimal policy $\pi^* = \{\pi_h^*\}_{h=1}^H$ is defined as [1]

$$\|V_1^* - V_1^\pi\|_\infty \coloneqq \sup_{x \in \mathcal{S}} V_1^*(x) - V_1^\pi(x) \;. \tag{2.4}$$

We describe the Bellman equation and the Bellman optimality equation for the evaluation setting. For any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the Bellman equation of the policy π takes the following form,

$$Q_h^\pi(x, a) = (r_h + \mathbb{P} V_{h+1}^\pi)(x, a) \;, \quad V_h^\pi(x) = \langle Q_h^\pi(x, \cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}} \;, \quad V_{H+1}^\pi(x) = 0 \;,$$

where $\langle Q_h^\pi(x, \cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}} = \int_{\mathcal{A}} Q_h^\pi(x, a) \pi_h(\mathrm{d}a \,|\, x)$ and $\mathbb{P}$ is the operator form of the transition kernel $\mathcal{P}_{F^*}$, i.e., defined as $(\mathbb{P} f)(x, a) = \mathbb{E}_{x' \sim \mathcal{P}_{F^*}(\cdot \,|\, x, a)}[f(x')]$ for any function $f : \mathcal{S} \to \mathbb{R}$. The subscript $\mathcal{A}$ is omitted subsequently if it is clear from the context. Similarly, the Bellman optimality equation takes the following form,

$$Q_h^*(x, a) = (r_h + \mathbb{P} V_{h+1}^*)(x, a) \;, \quad V_h^*(x) = \max_{a \in \mathcal{A}} Q_h^*(x, a) \;, \quad V_{H+1}^*(x) = 0 \;, \tag{2.5}$$

which implies that to find an optimal policy π^*, it suffices to estimate the optimal Q-function and then construct the greedy policy with respect to the optimal Q-function.

1. We should use esssup to be more measure-theoretically rigorous.

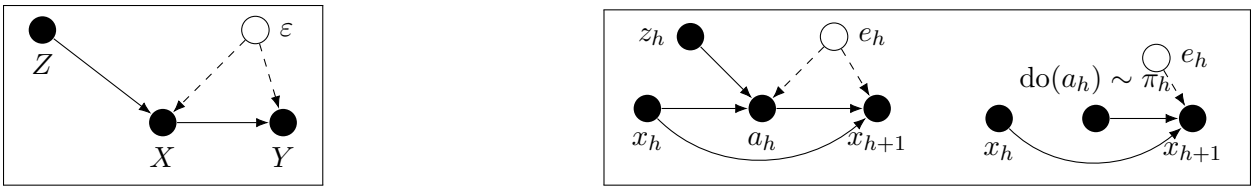


Figure 2: Left panel: An illustration of Definition 1 with one UC ε and three observable variables X, Y, and Z. Right panel: *Observation setting* of CMDP-IV with a behavior policy π_b (left). *Evaluation setting* of CMDP-IV with intervention induced by π (right).

2.2.2 Observation Setting: Data Collection Process

We describe the observation setting of CMDP-IV, in which we collect the data by executing the behavior policy $\pi_b \in \Delta(\mathcal{A}; \mathcal{S}, \mathcal{Z}, \mathcal{U})^H$. This distinguishes our work from most works in offline RL since we need to handle the issue of unobserved confounders, which makes the already difficult offline RL problem even more challenging.

At the beginning of each episode, the environment generates an initial state $x_1 \sim \xi_0$, a sequence of UCs $\{e_h\}_h \overset{\text{iid}}{\sim} \mathcal{P}_e$, and a sequence of observable IVs $\{z_h\}_h \overset{\text{iid}}{\sim} \mathcal{P}_z$. At the h-th step, given the current state x_h, the action a_h and the next state x_{h+1} are generated according to the following dynamics,

$$a_h \sim \pi_{b,h}(\cdot \,|\, x_h, z_h, e_h) \;, \quad x_{h+1} = F^*(x_h, a_h) + e_h \;. \tag{2.6}$$

The episode terminates if we reach the state x_{H+1} and we collect all observable variables, i.e., $\{(x_h, a_h, z_h, x'_h)\}_{h\in[H]}$, where $x'_h = x_{h+1}$ for any $h \in [H]$.

Assumption A.1. *The collection of random variables $\{e_1, \ldots, e_H, z_1, \ldots, z_H, x_1\}$ are independent. Moreover, we assume the marginal distribution of confounders are identical, and that and the marginal distribution of confounders of instruments are identical, i.e., $e_h \sim \mathcal{P}_e$, $z_h \sim \mathcal{P}_z$ for all $h \in [H]$.*

A causal DAG is given in Figure 2 (right) to graphically illustrate such dynamics. At any stage h, the variable z_h is an IV relative to the pair (a_h, x_{h+1}). Indeed, z_h affects the action a_h only through Eq. (2.6), and its effect on x_{h+1} must be channelled through a_h because it does not appear in the second equation in Eq. (2.6).

The main difference between the evaluation setting Eq. (2.1) and the observation setting Eq. (2.6) is whether the UC e_h has an effect on the action a_h. In the language of causal inference (Pearl, 2009), a policy $\pi = \{\pi_h \in \Delta(\mathcal{A}; \mathcal{S})\}_{h=1}^H$ induces the stochastic intervention $\mathrm{do}(a_1 \sim \pi_1(\cdot \,|\, x_1), \ldots, a_H \sim \pi_H(\cdot \,|\, x_H))$ on the DAG in Figure 2 (left part of the right panel), and the resulting DAG is obtained by removing all arrows pointing into the action a_h; see Figure 2 (right part of the right panel). Appendix B includes more details on the do-operation.

Under the observation and the evaluation settings described in Sections 2.2.2 and 2.2.1, respectively, we aim to answer the following question:

Given data collected from the confounded dynamics Eq. (2.6) *in the observation setting, can we find a policy that minimizes the suboptimality defined in Eq.* (2.4) *in the evaluation setting?*

We now remark on the modeling assumptions.

Remark 2 (Generalization of Figure 2 (right panel)). *We have made two simplifying assumptions. First, we assume e_h only confounds the transition dynamics (the arrow from a_h to x_{h+1}). The unobservables e_h could also affect the action and the reward, or state and reward, or both. Second, we assume in each stage, z_h and e_h are generated in an i.i.d. manner and are independent of all other random variables in the MDP. In practice it is likely that the sequences $\{z_h\}$ and $\{e_h\}$ exhibit temporal dependence. We focus on this simplified model because it captures the essence of IVs: a variable that affects x_{h+1} only through the action a_h. In the work of Bennett et al. (2021) where the authors study policy evaluation with unobserved confounders, confounders are also assumed i.i.d.*

Remark 3 (On additive noise assumption). *A more general version of this problem, which we leave for future work, would be the setting where the transition dynamics are of the form $x_{t+1} = F(x_h, a_h, e_h)$, in contrast to our additive Gaussian noise assumption. We remark non-identification is a key issue in the fully non-parametric model. Let us revisit the IV diagram presented in Figure 2, which represents the simplest case of an IV with structural equations $Y = f(X, \varepsilon)$ and $X = g(Z, \varepsilon)$, with $Z \perp\!\!\!\perp \varepsilon$. It is well-known that the conditional independence implied by the IV diagram is not enough to identify the causal effect of X on Y (Bareinboim and Pearl, 2012; Hünermund and Bareinboim, 2023). Roughly this means there exist two distributions of random variables (X, Y, Z) that are compatible with the IV diagram, and yet the structural functions f are different. One could instead work with a partially identified IV model, using bounds of the causal effects (Balke and Pearl, 1994, 1997; Zhang and Bareinboim, 2021).*

Remark 4 (The challenge of UCs). *The challenge stems from the fact that the UC e_h enters both of the equations Eq. (2.6). For ease of discussion, suppose that the behavior policy π_b is deterministic. With slight abuse of notations, we denote by $\pi_{b,h} : \mathcal{S} \times \mathcal{Z} \times \mathcal{U} \to \mathcal{A}$ the deterministic behavior policy at the h-th step for any $h \in [H]$. Now, Eq. (2.6) writes $a_h = \pi_{b,h}(x_h, z_h, e_h)$. We further assume that the behavior policy $\pi_{b,h}(x, z, e)$ is invertible in the third argument e for any $(x, z) \in \mathcal{S} \times \mathcal{Z}$, which allows us to define its inverse $\pi_{b,h}^{-1} : \mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to \mathcal{U}$. Then, by substituting $e_h = \pi_{b,h}^{-1}(x_h, z_h, a_h)$ into Eq. (2.6), we have $x_{h+1} = F^*(x_h, a_h) + \pi_{b,h}^{-1}(x_h, z_h, a_h)$. By taking expectation conditioning on (x_h, a_h), we obtain $\mathbb{E}[x_{h+1} \,|\, x_h, a_h] = F^*(x_h, a_h) + \delta(x_h, a_h)$, where $\delta(x_h, a_h) := \mathbb{E}[\pi_{b,h}^{-1}(x_h, z_h, a_h) \,|\, x_h, a_h]$. This indicates that the true transition function F^* cannot be obtained by simply regressing x_{h+1} on (x_h, a_h), since that would result in a biased estimate.*

Remark 5 (Global IVs and global UCs). *Our method directly extends to cases where, instead of a time-varying IV, we only have access to a global IV that affects all the actions taken on a trajectory simultaneously, e.g. a doctor's preference to certain treatments. The reason is that the global IV, conditional on the past history, is also a valid IV for each time step, mimicking the structure of the time-varying IV. Specifically, having a global IV is equivalent to having $z_h = z$ for all h, i.e. all local IVs take the same value. Then, by the full independence between $\{e_h\}_h$ and z, the core requirement of the time-varying IV $\mathbb{E}[e_h \,|\, z] = 0$ still holds, and thus our result applies.*

We feel that policy learning would be difficult if the global confounder affects both actions and states. In more detail, consider global UCs that affect all stages of decision making, and thus affect all states x_h and actions a_h. While IV can deconfound the effects of global UCs on the actions a_h, it cannot deconfound their effects on the states x_h, x_{h+1}. The transition dynamics from x_h to x_{h+1} would depend on the global UCs. This dependence would limit the performance of the learned policy in evaluation settings if the evaluation transition dynamics from x_h to x_{h+1} does not depend on the global UCs in the same way.

Global confounders do not seem a natural extension in our additive dynamics. For example, suppose the dynamics for stage h write

$$a_h \sim \pi_b(\cdot \,|\, x_h, z_h, e) \,, \quad x_{h+1} = F^*(x_h, a_h) + e \,,$$

where the UC at each stage is identical and is denoted e. One can difference the sequence $\{x_h\}_h$, and obtain $x_{h+1} - x_h = F^*(x_h, a_h) - F^*(x_{h-1}, a_{h-1})$, where the global UC is cancelled. Due to these considerations, we focus on the CMDP-IV setting in this work, which itself is a natural extension of the IV model introduced in Section 2.1 to the multi-stage decision making process.

3. IV-Aided Value Iteration

How can an IV help us design an offline RL algorithm? To answer this question, we proceed by a model-based approach. We estimate the transition function F^* first. And then any planning algorithm (value iteration in our case) can be used to recover the optimal policy under the evaluation setting.

3.1 A Primal-Dual Estimand

We observe that, thanks to the presence of IVs, the transition function F^* is the solution of a conditional moment restriction (CMR). To estimate the transition function F^* based on the CMR, we derive a primal-dual formulation of the CMR in Section 3.1.2.

3.1.1 Conditional Moment Restriction

Following the confounded dynamics Eq. (2.6), the behavior policy π_b induces the distribution of the observable trajectories $\{x_h, a_h, z_h, x'_h = x_{h+1}\}_{h=1}^H$. We denote by d_{h,π_b} the distribution of the tuple $(x_h, a_h, z_h, x'_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z} \times \mathcal{S}$ at the h-th step for any $h \in [H]$, i.e., $d_{h,\pi_b}(x, a, z, x')$. We further define the *average visitation distribution* as follows,

$$\bar{d}_{\pi_b}(x, a, z, x') := \frac{1}{H} \cdot \sum_{h=1}^{H} d_{h,\pi_b}(x, a, z, x') \tag{3.1}$$

for any $(x, a, z, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z} \times \mathcal{S}$. We denote by $L^2(\mathcal{S}, \mathcal{A}) = \{f \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R},\ \mathbb{E}[f(x, a)^2] < \infty\}$ the space of square integrable functions equipped with the norm $\|f\|^2_{L^2(\mathcal{S},\mathcal{A})} = \mathbb{E}[f(x, a)^2]$. Similarly, we define $L^2(\mathcal{Z})$ and the norm $\|g\|^2_{L^2(\mathcal{Z})} = \mathbb{E}[g(z)^2]$. The operator $\mathcal{T} \colon L^2(\mathcal{S}, \mathcal{A}) \to L^2(\mathcal{Z})$ is defined as

$$(\mathcal{T} f)(\cdot) = \mathbb{E}[f(x, a) \,|\, z = \cdot\,]\,. \tag{3.2}$$

The following proposition states the conditional moment restriction (CMR) implied by the IVs in the observational confounder dynamics Eq. (2.6). See Appendix C.1 for the proof.

Proposition 6 (CMR). *If (x, a, z, x') is distributed according to the law $\bar{d}_{\pi_b}$, then for any $z \in \mathcal{Z}$,*

$$\mathbb{E}[F^*(x, a) \,|\, z] = \mathbb{E}[x' \,|\, z]\,. \tag{3.3}$$

Proposition 6 implies that the transition function F^* satisfies the equation $\mathcal{T} F^* = \mathbb{E}[x' \,|\, z]$, where the operator $\mathcal{T}$ is defined in Eq. (3.2). Such an equation is a Fredholm integral equation of the first kind (Kress, 1989). Given data collected from $\bar{d}_{\pi_b}$, we aim to estimate F^* based on the CMR.

3.1.2 A Primal-Dual Estimand

We derive a primal-dual estimand for $F^* = [f_1^*, \ldots, f_{d_x}^*]^\top$. For any $i \in [d_x]$, by Proposition 6, $\mathbb{E}[f_i^*(x, a) \,|\, z] = \mathbb{E}[x'_i \,|\, z]$, where x'_i is the i-th element of the next state x'. We find f_i^* by solving the

least-square problem $\min_{f_i \in L^2(\mathcal{S},\mathcal{A})} \frac{1}{2}\mathbb{E}[(\mathbb{E}[f_i(x,a) \mid z] - \mathbb{E}[x_i' \mid z])^2]$. By Fenchel duality, the least-square problem admits a primal-dual formulation

$$\min_{f_i \in L^2(\mathcal{S},\mathcal{A})} \max_{u_i \in L^2(\mathcal{Z})} \left\{ \mathbb{E}\big[(f_i(x,a) - x_i')u_i(z)\big] - \tfrac{1}{2}\mathbb{E}\big[u_i(z)^2\big] \right\}, \tag{3.4}$$

where u_i is the dual variable. To approximate the L^2 spaces, we introduce two known feature maps

$$\phi \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_\phi}, \quad \psi \colon \mathcal{Z} \to \mathbb{R}^{d_\psi},$$

and let $\mathcal{H}_\phi$ and $\mathcal{H}_\psi$ denote the spaces spanned by ϕ and ψ, respectively. For simplicity, we define the following uncentered covariance matrices

$$A := \mathbb{E}[\psi(z)\phi(x,a)^\top], B := \mathbb{E}[\psi(z)\psi(z)^\top], C := \mathbb{E}[x'\psi(z)^\top], D := \mathbb{E}[\phi(x,a)\phi(x,a)^\top]. \tag{3.5}$$

where the expectations are taken following $\bar{d}_{\pi_b}$. We replace the L^2 spaces in Eq. (3.4) by their finite-dimensional subspaces,

$$\min_{f_i \in \mathcal{H}_\phi} \max_{u_i \in \mathcal{H}_\psi} \left\{ \mathbb{E}[(f_i(x,a) - x_i')u_i(z)] - \tfrac{1}{2}\mathbb{E}[u_i(z)^2] \right\},$$

which, in matrix form, writes

$$\min_{\theta_i \in \mathbb{R}^{d_\phi}} \max_{\omega_i \in \mathbb{R}^{d_\psi}} \left\{ \omega_i^\top A\theta_i - b_i^\top \omega_i - \tfrac{1}{2}\omega_i^\top B\omega_i \right\}, \tag{3.6}$$

where $b_i := \mathbb{E}[x_i'\psi(z)]$ and A and B are defined in Eq. (3.5). We address the approximation error incurred by such finite-dimensional approximation in Section 4.2. Now we collect Eq. (3.6) for all coordinates $i \in [d_x]$, giving the key primal-dual estimand W^{sad}

$$W^{\text{sad}} := \operatorname*{argmin}_W \max_K L(W,K), \tag{3.7}$$

where $L(W,K) := \text{Tr}(KAW^\top) - \text{Tr}(CK^\top) - \frac{1}{2}\text{Tr}(KBK^\top)$ with $W = [\theta_1, \ldots, \theta_{d_x}]^\top \in \mathbb{R}^{d_x \times d_\phi}$ and $K = [\omega_1, \ldots, \omega_{d_x}]^\top \in \mathbb{R}^{d_x \times d_\psi}$. For appropriately chosen feature maps we expect $W^{\text{sad}}\phi \approx F^*$.

3.2 Algorithm

We first introduce the following data sampling assumption for the algorithm.

Assumption A.2 (Observation data)**.** *We have access to i.i.d. data from the average visitation distribution defined in Eq.* (3.1)*. That is,* $\{(x_t, a_t, z_t, x_t')\}_{t=0}^{T-1} \overset{\text{iid}}{\sim} \bar{d}_{\pi_b}$.

Assumption A.2 is only used to simplify the presentation of our results, by ignoring the temporal dependence in the data.

Algorithm 1 introduces the backbone of the paper, IV-aided Value Iteration (IVVI), which recovers the optimal policy under the evaluation setting given data collected from the confounded dynamics under the observation setting. Algorithm 1 consists of the following two phases.

Phase 1. In Lines 3–7 of Algorithm 1, we solve Eq. (3.7) using stochastic gradient descent-ascent. At the t-th iteration, we have $\frac{\partial L}{\partial W} = K_t A$, $\frac{\partial L}{\partial K} = -(K_t B + C - W_t A^\top)$, which combined with the definitions of A, B, and C in Eq. (3.5), gives us the updates of W_{t+1} and K_{t+1} in Line 5, respectively.

Algorithm 1 IV-aided Value Iteration (IVVI)

1: **Input:** Reward functions $\{r_h\}_{h=1}^H$, feature maps ϕ and ψ, iterations T, stepsizes $\{\eta_t^\theta, \eta_t^\omega\}_{t=1}^T$, initial estimates K_0 and W_0, variance σ^2, samples $\{(x_t, a_t, z_t, x_t')\}_{t=0}^{T-1}$ in Assumption A.2.
2: **Phase 1 (Estimation of W^{sad} in Eq. 3.7)**
3: **for** $t = 0, 1, \ldots, T-1$ **do**
4: $\quad \phi_t \leftarrow \phi(x_t, a_t)$, $\psi_t \leftarrow \psi(z_t)$.
5: $\quad W_{t+1} \leftarrow W_t - \eta_t^\theta \cdot (K_t \psi_t \phi_t^\top)$, $\quad K_{t+1} \leftarrow K_t + \eta_t^\omega \cdot (K_t \psi_t \psi_t^\top + x_t' \psi_t^\top - W_t \phi_t \psi_t^\top)$.
6: **end for**
7: **Phase 2 (Value iteration)**
8: $\widehat{V}_{H+1}(\cdot) \leftarrow 0$, $\quad \widehat{W} \leftarrow W_T$.
9: **for** $h = H, H-1, \ldots, 1$ **do**
10: $\quad \widehat{Q}_h(\cdot, \cdot) \leftarrow r_h(\cdot, \cdot) + \int_{\mathcal{S}} \widehat{V}_{h+1}(x') \mathcal{P}_{\widehat{W}}(\mathrm{d}x' \,|\, \cdot, \cdot)$.
11: $\quad \widehat{\pi}_h(\cdot) \leftarrow \mathrm{argmax}_a \, \widehat{Q}_h(\cdot, a), \quad \widehat{V}_h(\cdot) \leftarrow \max_a \widehat{Q}_h(\cdot, a)$.
12: **end for**
13: **Output:** $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$.

Phase 2. Given the estimated matrix $\widehat{W}$ generated from Phase 1, in Lines 8–12 of Algorithm 1, we implement value iteration to recover an optimal policy for the evaluation setting. In the optimality Bellman equation Eq. (2.5), we replace the true transition operator $\mathbb{P}$ with the estimated transition operator induced by $\widehat{W}$, i.e., $\widehat{Q}_h(x, a) = r_h(x, a) + (\widehat{\mathbb{P}}\widehat{V}_{h+1})(x, a)$, for any $(x, a) \in \mathcal{S} \times \mathcal{A}$. Here, $\widehat{\mathbb{P}}$ is the operator form of $\mathcal{P}_{\widehat{W}} := \mathcal{P}_{\widehat{W}\phi}$, such that $(\widehat{\mathbb{P}} f)(x, a) = \mathbb{E}_{x' \sim \mathcal{P}_{\widehat{W}}(\cdot \,|\, x, a)}[f(x')]$ for any $f \colon \mathcal{S} \to \mathbb{R}$.

We remark that to efficiently implement the integration and maximization in Phase 2 of Algorithm 1, one can use Monte Carlo integration and gradient methods, respectively.

4. Theory

We first introduce two assumptions on the feature maps ϕ and ψ.

Assumption A.3 (Bounded feature maps)**.** *We have* $\|\phi(x, a)\|_2 \le 1$ *and* $\|\psi(z)\|_2 \le 1$ *for any* $(x, a, z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z}$.

Assumption A.4 (Nondegenerate feature maps)**.** *It holds that* $\mathrm{rank}(A) = d_\phi$ *and* $\mathrm{rank}(B) = d_\psi$ *for* A *and* B *defined in Eq.* (3.5).

Uniqueness of W^{sad}. Assumption A.4 implies the minimax problem Eq. (3.7) admits a unique solution. In the min-max problem Eq. (3.7), for a fixed primal variable W, the unique maximizer $K^*(W)$ of the inner problem in takes the form $K^*(W) := (WA^\top - C)B^{-1}$. This holds by the invertibility of B, whose minimum eigenvalue is now denoted by $\mu_B := \sigma_{\min}(B) > 0$. Plug in this optimal value we have $\max_K L(W, K) = \frac{1}{2}\mathrm{Tr}[(WA^\top - C)B^{-1}(WA^\top - C)^\top]$. By full-rankness of A we know W^{sad} is the unique minimizer of the map $W \mapsto \max_K L(W, K)$.

Instrument Strength. Assumption A.4 implicitly impose sufficient correlation between $\phi(x, a)$ and $\psi(z)$. In other words, IVs needs to be strong to have enough explanatory power for the behavior policy π_b. Weak IV is a well-known pitfall in applied economic research (Angrist and Pischke, 2008). For RL applications with confounded data, practitioners should take into account domain knowledge of the behavior policy to avoid using weak IVs. We introduce a quantity μ_{IV},

which quantifies the strength of IVs. We define the IV strength μ_{IV} as follows,

$$\mu_{\mathrm{IV}} := \inf \left\{ \frac{\|\Pi_\psi \mathcal{T} f\|^2_{L^2(\mathcal{Z})}}{\|f\|^2_\phi} \,\middle|\, f \in \mathcal{H}_\phi, \|f\|_\phi \neq 0 \right\}, \tag{4.1}$$

where Π_ψ is the projection operator onto the space $\mathcal{H}_\psi$, i.e., $\Pi_\psi u = \mathrm{argmin}_{u' \in \mathcal{H}_\psi} \|u - u'\|^2_{L^2(\mathcal{Z})}$ for any $u \in L^2(\mathcal{Z})$. The definition of μ_{IV} in Eq. (4.1) mimics the notion of *sieve measure of ill-posedness* well-known in the literature on NPIV as a measure of IV strength (Blundell et al., 2007; Chen and Christensen, 2018). We next show μ_{IV} admits a simple expression.

Proposition 7. *Let A.4 hold. Then* $\mu_{\mathrm{IV}} = \sigma_{\min}(A^\top B^{-1} A)$.

4.1 Parametric Case

We impose the following assumptions on the transition function F^* and the conditional expectation operator $\mathcal{T}$.

Assumption A.5 (Linear representation)**.** *It holds* $F^* = W^* \phi$ *for some* $W^* \in \mathbb{R}^{d_x \times d_\phi}$.

Such a linear form of the transition function F^* is commonly assumed in the literature (Kakade et al., 2020; Mania et al., 2022) in the context of dynamical system identification.

Assumption A.6 (Realizability)**.** *For all* $f \in \mathcal{H}_\phi$, *it holds that* $\mathcal{T} f \in \mathcal{H}_\psi$.

Proposition 8. *Let A.4, A.5 and A.6 hold. Then* $W^* = W^{\mathrm{sad}}$.

One important contribution of our work is that we quantify how the strength of the IV is playing a role in terms of recovering optimal policy from confounded data. We provide a sketch of the proof for Theorem 9 in Appendix A. The complete proofs are given in Appendix C.4.

Theorem 9 (Parametric case)**.** *Let A.2–A.6 hold. There exists a choice for stepsizes in Algorithm 1 of the form* $\eta^\theta_t = \beta/(\gamma + t)$ *and* $\eta^\omega_t = \alpha \eta^\theta_t$ *for any* $t \in [T]$, *where* $\alpha = c_1 \mu_{\mathrm{IV}}^{-1} \mu_B^{-1.5}$, $\beta = c_2 \mu_{\mathrm{IV}}^{-1}$, $\gamma = c_3 \mu_{\mathrm{IV}}^{-4} \mu_B^{-3.5}$, *and* c_1, c_2, c_3 *are positive absolute constants, such that*

(i) the estimation error satisfies

$$\mathbb{E}\big[\|W_T - W^*\|_F^2\big] \leq \frac{\nu}{\gamma + T}, \tag{4.2}$$

where $\nu = \max\{\gamma \widetilde{P}_0, c_4 \mu_{\mathrm{IV}}^{-4} \mu_B^{-2.5} \cdot d_x \sigma^2\}$ *and* $\widetilde{P}_0 = \|W_0 - W^*\|_F^2 + \sqrt{\mu_B} \cdot \|K_0 - K^*(W_0)\|_F^2$ *with* c_4 *being a positive absolute constant. And*

(ii) the planning error satisfies

$$\mathbb{E}\big[\|V_1^* - V_1^{\widehat{\pi}}\|_\infty\big] \leq H \cdot \min\left\{ 2H\sigma^{-1} \sqrt{\frac{\nu}{\gamma + T}}, 1 \right\}. \tag{4.3}$$

The expectation is taken over the data.

For an appropriately chosen initial estimates W_0 and K_0, Theorem 9 shows that the sample complexity needed to recover an ϵ-optimal policy using observational data is of order

$$O(\mu_{\mathrm{IV}}^{-4} \mu_B^{-2.5} \cdot H^4 d_x \sigma^2 \epsilon^{-2}),$$

where μ_{IV} characterizes IV strength, i.e., how well the IV is able to explain the behavior policy, μ_B quantifies the compatibility of the dual feature map and the IV, H is the horizon of the MDP, and d_x is the dimension of states. To the best of our knowledge, this is the first sample complexity result for recovering optimal policy using confounded data when a valid IV is present.

Remark 10 (Joint computational and statistical efficiency). *The estimation procedure (phase 1) is readily a scalable algorithm, in contrast to estimators defined as the saddle-point of a finite-sum; see Remark 13. From an optimization perspective, the saddle-point problem Eq.* (3.6) *is a stochastic convex-strongly-concave one, a case rarely investigated in the optimization literature. The asymmetric structure in the primal and dual variables demands more detailed analysis of the algorithm in order to achieve a fast $O(1/T)$ rate.*

We now review literature that studies convex-strongly-concave (CSC) stochastic saddle point problem. A slow rate $O(1/\sqrt{T})$ is obvious by the results for general stochastic convex-concave problem (Nemirovski et al., 2009). The work of Chambolle and Pock (2011) studies deterministic CSC problem with bilinear coupling and shows a $O(1/T^2)$ rate. Wang and Xiao (2017); Du et al. (2017); Du and Hu (2019) consider CSC problem with finite sum structure and bilinear coupling structure, and shows a linear convergence rate by variance reduction techniques. In contrast, our algorithm solves stochastic CSC problem with linear coupling structure with a fast $O(1/T)$ rate without the need of projection. Moreover, the assumption of bounded variance of the stochastic gradient does not hold in our case, rendering most existing analysis invalid.

Remark 11 (Dependence on IV strength). *In Eq.* (4.3)*, for appropriately chosen initial estimates W_0 and K_0, only the second term in the definition of ν matters. We are effectively solving d_x NPIV problems, and the asymptotic order for solving just one NPIV problem is $O(\mu_{\mathrm{IV}}^{-4}\mu_B^{-2.5}\sigma^2 T^{-1})$. The dependence on the dimension of feature maps d_ϕ and d_ψ is hidden in the minimum eigenvalues μ_B and μ_{IV}. We compare our result with the work by Dikkala et al. (2020) under A.6. There the proposed estimator is the saddle-point of the sample version of Eq.* (3.6)*; see Remark 13 for more details. In particular, they provide a bound in the L^2-norm, and the order of the variance term is $O(\mu_{\mathrm{IV}}^{-4}\max\{d_\phi, d_\psi\}T^{-1})$ [2]. The minimax optimal rate for NPIV problem is established in the work of Blundell et al. (2007), attained by sieve estimators. In comparison, the variance term in the minimax optimal rate is of order $O\big(\widetilde{\mu_{\mathrm{IV}}}^{-2} d_\psi T^{-1}\big)$ [3], where $\widetilde{\mu_{\mathrm{IV}}}$ is the minimum nonzero singular value of $D^{-1/2}AB^{-1/2}$, quantifying the strength of an IV in a similar way to our μ_{IV}.*

Remark 12 (Dependence on horizon and state dimension). *The work of Kakade et al. (2020) provides a $\sqrt{T}$-regret bound for online learning of an additive nonlinear dynamics. Their regret bound translates to a $O(d_\phi(d_\phi + d_x + H)H^3\epsilon^{-2})$ sample complexity bound, ignoring logarithmic factors; see Corollary 3.3 of Kakade et al. (2020). Despite that we deal with confounders in additive nonlinear dynamics, our dependence on d_x and H matches their sample complexity bounds.*

Remark 13 (Stochastic approximation for instrumental variables). *Our stochastic approximation (SA) estimation procedure is in contrast with the empirical saddle-point estimator proposed in Dikkala et al. (2020). To estimate f_j^*, their estimator would be defined as the solution to the finite-sum saddle-point problem*

$$\operatorname*{argmin}_{f\in\mathcal{H}_\phi}\max_{u\in\mathcal{H}_\psi}\frac{1}{n}\sum_{i=1}^{n}\Big\{(f(x_i,a_i)-x'_{i,j})u(z_i)+\frac{1}{2}u(z_i)^2\Big\}-\frac{\lambda}{2}\|u\|^2_{\mathcal{H}_\phi}+\frac{\mu}{2}\|f\|^2_{\mathcal{H}_\psi} \tag{4.4}$$

for some positive λ and μ. Here the data $\{x_i, a_i, z_i, x'_i\}$ are i.i.d. draws from $\bar{d}_{\pi_b}$, and $x'_{i,j}$ denotes the j-th coordinate of $x'_i \in \mathbb{R}^{d_x}$. Their procedure faces two challenges: (i) choosing the correct regularization parameter, and (ii) finding an approximate solution of the convex-concave optimization problem Eq. (4.4)*, which requires a separate discussion of computational complexity. The theoretical trade-off among regularization bias, statistical error and optimization error is unclear, as is shown*

2. In Appendix D of Dikkala et al. (2020), their (γ_n, k_n, m_n) is the same as our $(\mu_{\mathrm{IV}}^{-1}, d_\phi, d_\psi)$.
3. In Theorem 2 of Blundell et al. (2007), their (τ_n, k_n) is the same as our $(\widetilde{\mu_{\mathrm{IV}}}^{-1}, d_\phi)$.

in related primal-dual methods in RL; see, e.g., Dai et al. (2017, 2018); Nachum et al. (2019). In contrast, the SA approach considered in this work tackles computational error and statistical error jointly and enjoys a fast rate of $O(1/T)$.

4.2 Nonparametric Case

In A.5 and A.6 we make the simplifying assumption that both the true transition function F^* and the image of the operator $\mathcal{T}$ lie in some known finite dimensional spaces. To extend out theory to the nonparametric case (e.g., $F^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_x}$ is Hölder continuous, and functions of the form $\{\mathcal{T}f \,|\, f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, bounded and continuous$\}$ are also Hölder continuous), we need to discuss two issues.

The first one is identification: whether F^* is the unique solution to the CMR Eq. (3.3). Identification in NPIV usually requires some form of completeness assumptions. For example, bounded completeness condition is a relatively weak regularity assumption on the average visitation distribution $\bar{d}_{\pi_b}$. For two random variables X and Y, X is boundedly complete w.r.t Y if for all Y-a.s. bounded function f, it holds $\mathbb{E}[f(Y) \,|\, X] = 0$ implies $f = 0$ Y-a.s. Intuitively, it requires that the distribution of Y exhibits a sufficient amount of variation when conditioning on different values of X. It is well-known that there is a wide range of distributions that satisfy bounded completeness; see, for example, Blundell et al. (2007); D'Haultfoeuille (2011); Hu and Shiu (2017); Andrews (2017).

The second issue is the error caused by finite-dimensional approximation which we address below. Let f^* be one element of $F^* = [f_1^*, \ldots, f_{d_x}^*]^\top$. If A.5 is violated, we define the primal approximation error

$$\eta_1 := \|f^* - \Pi_\phi f^*\|_{L^2(\mathcal{S},\mathcal{A})}.$$

If A.6 is violated, we define the dual error, which characterizes how well the dual function space $\mathcal{H}_\psi$ approximates functions of the form $\mathcal{T}(f - f^*)$ for $f \in \mathcal{H}_\phi$. Formally we define

$$\eta_2 := \sup\{\|\mathcal{T}f - \Pi_\psi \mathcal{T} f\|_{L^2(\mathcal{Z})} : f \in \mathcal{H}_\phi, \|f\|_{L^2(\mathcal{S},\mathcal{A})} \le 1\}.$$

Obviously A.5 implies $\eta_1 = 0$ and A.6 implies $\eta_2 = 0$.

We show that, when A.5 and A.6 are violated, the difference between F^* and $W^{\mathrm{sad}}\phi$ has only linear dependence on the approximation errors η_1 and η_2. Notably, the dual error η_2 is inflated by μ_{IV}^{-1}. Recall f_i^* is the i-th element of $F^* = [f_1^*, \ldots, f_{d_x}^*]^\top$, and W_i^{sad} is the i-th row of the estimand W^{sad} defined in Eq. (3.7).

Theorem 14 (Nonparametric case)**.** *Let A.2–A.4 hold. Assume there is a constant $c > 0$ such that $\mu_{\mathrm{IV}}^{-1} \cdot \|\mathcal{T}(f_i^* - \Pi_\phi f_i^*)\|_{L^2(\mathcal{Z})} \le c \cdot \|f_i^* - \Pi_\phi f_i^*\|_{L^2(\mathcal{S},\mathcal{A})}$. We define the operator $Q \colon L^2(\mathcal{S},\mathcal{A}) \mapsto \mathcal{H}_\phi$, $Qf = \mathrm{argmin}_{f' \in \mathcal{H}_\phi} \|\Pi_\psi \mathcal{T}(f' - f)\|_{L^2(\mathcal{Z})}$. Let $\mu = \|(\Pi_\phi - Q) f_i^*\|_{L^2(\mathcal{S},\mathcal{A})}$. It holds*

$$\|f_i^* - W_i^{\mathrm{sad}} \cdot \phi\|_{L^2(\mathcal{S},\mathcal{A})} \le (1 + 2c) \cdot \eta_1 + \mu_{\mathrm{IV}}^{-1} \cdot \mu \cdot \eta_2 \,.$$

In Theorem 14, the existence of such a constant c is called the stability assumption; see Blundell et al. (2007) and Chen and Christensen (2018) for a detailed discussion. Note the dual approximation error η_2 is inflated by a factor of μ_{IV}^{-1}.

The estimation phase still produces an estimator that converges to W^{sad} at $O(1/T)$ rate. The only difference is, in the planning phase, we are performing value iteration with a biased model.

5. Experiment

In this section, we present numerical experiments on the parametric and nonparametric cases described in Section 4. The goal of this section is to verify that Algorithm 1 successfully identifies

the transition model based on sequential observational data and recovers the optimal policy by planning with the estimated transition model. Importantly, we aim to quantify how the strength of instrument affects estimation of causal quantities in the sequential setting.

All experiments in this section can be reproduced with the code at `https://github.com/ChampionRecLuse/ivvi`.

5.1 Parametric Setting

We consider data generation procedures with a linear transition dynamic. We use our algorithm in 5-dimension, 10-dimension, and 20-dimension. In order to compare results in different dimensions, we use similar parameter setting in different dimensions. To be specific, we let k stands for the number of dimensions. The CMDP-IV operates in the following spaces: $\mathcal{S} = \mathcal{U} = \mathcal{Z} = \mathbb{R}^k$, $\mathcal{A} = [-1, 1]^k$. We generate 80 episodes, where each episode has a horizon $H = 1000$. The environment starts from an all-ten vector x_0, for example, $x_0 = [10, 10, 10, 10, 10]^\top$ when $k = 5$. The environment generates a sequence of UCs $\{e_h\}_h \overset{\text{iid}}{\sim} \mathcal{P}_e$ and a sequence of observable IVs $\{z_h\}_h \overset{\text{iid}}{\sim} \mathcal{P}_z$. We let e_h and z_h follow Gaussian distributions, i.e., $\mathcal{P}_e = \mathcal{N}(\mu_e, \Sigma_e)$, $\mathcal{P}_z = \mathcal{N}(\mu_z, \Sigma_z)$. At the h-th step, the next state x_{h+1} is generated by the equations

$$a_h \sim \text{Proj}_{[-1,1]^k}\left(\mathcal{N}(z_h + e_h, \Sigma_a)\right), \quad x_{h+1} = F^*(x_h, a_h) + e_h = Px_h - Qa_h + e_h,$$

where

$$\Sigma_a = \begin{pmatrix} 1 & 0.3 & 0 & \dots & 0 \\ 0.3 & 1 & 0.3 & \ddots & \vdots \\ 0 & 0.3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & 0.3 \\ 0 & \dots & 0 & 0.3 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 0.5 & 0.2 & 0 & \dots & 0 \\ 0.2 & 0.5 & 0.2 & \ddots & \vdots \\ 0 & 0.2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0.5 & 0.2 \\ 0 & \dots & 0 & 0.2 & 0.5 \end{pmatrix},$$

$$Q = \begin{pmatrix} 0.5 & 0.1 & 0 & \dots & 0 \\ 0.1 & 0.5 & 0.1 & \ddots & \vdots \\ 0 & 0.2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0.5 & 0.1 \\ 0 & \dots & 0 & 0.1 & 0.5 \end{pmatrix}.$$

Note that each matrix above has size $k \times k$. Here a_h is generated by first sampling from a Gaussian and then projecting it onto the cube $[-1, 1]^k$ w.r.t. the Euclidean distance. The projection here is mainly to stabilize the state dynamics of the behavior policy. The reward function is $r_h(x_h, a_h) = -0.05\|x_h\|^2$. We can tell the optimal policy easily under this reward. The optimal policy forces the agent to get close to zero vector since this state has the highest reward. We use different covariance matrices Σ_e, Σ_z to control the instrument strength (defined in Section 4) and study its influences on our algorithm. The intuition is that if the Gaussian distribution of UCs is steeper or the Gaussian distribution of IV is flatter, we can have a stronger instrument strength. In each dimension, we consider three data generation processes, all of which correspond to an identical transition function. In the experiments, we use the following pairwise covariance matrices:

I) $\Sigma_z = 0.7I_k$ and $\Sigma_e = \Sigma_a$; II) $\Sigma_z = I_k$ and $\Sigma_e = \Sigma_a$; III) $\Sigma_z = 1.5I_k$ and $\Sigma_e = \Sigma_a$.

In the above setup, it is obvious that z_h is a valid IV and e_h is a valid UC in the environment.

In order to check the robustness of the proposed method, we introduce variations in the covariance matrices Σ_a and Σ_e, corresponding to the action generation process and error term, respectively. This also helps to check the robustness of our method when the transition kernel is multivariate normal with general covariance matrices. We construct symmetric Toeplitz matrices where each sub-diagonal has constant values that decay exponentially as the covariance matrices. A symmetric Toeplitz matrix $T(\rho) \in \mathbb{R}^{k\times k}$ with parameter ρ is defined as $T_{ij} = \rho^{|i-j|}$, where T_{ij} is the i, j element of $T(\rho)$ and ρ is the decay parameter. In the experiment, we use the following pairwise covariance matrices:

$$\begin{aligned}
\text{IV)}\ & \Sigma_z = 0.7I_k, \Sigma_e = T(-0.5) \text{ and } \Sigma_a = T(-0.3);\\
\text{V)}\ & \Sigma_z = I_k, \Sigma_e = T(0.1) \text{ and } \Sigma_a = T(0);\\
\text{VI)}\ & \Sigma_z = 1.5I_k, \Sigma_e = T(0.5) \text{ and } \Sigma_a = T(0.3).
\end{aligned}$$

It is important to note that we collect data based on these settings and use our method to estimate the transition function. The training and testing environments are consistent with settings I, II, and III.

We instantiate Algorithm 1 with feature maps $\phi(x, a) = [1, x, a]^\top$ and $\psi(x, z) = [1, x, z]^\top$. Note that the current state x is also a variable in the feature map ψ. According to the feature map, the true transition function can be written as $W^*\phi(x, a)$, where $W^* = \begin{bmatrix}\mathbf{0} & P & -Q\end{bmatrix}$. We use the minibatch stochastic gradient descent ascent to compute W^{sad}. We set initial estimates K_0 as a zero matrix and W_0 as a matrix where every entry is one fifth. For 5-dimension and 10-dimension case, at t-th iteration, stepsizes we use are $\eta_t^\theta = 0.05 + \frac{1}{18+t}$ and $\eta_t^\omega = \frac{1}{18+t}$. For 20-dimension case, stepsizes are $\eta_t^\theta = 0.06 + \frac{1}{18+t}$ and $\eta_t^\omega = \frac{1}{18+t}$. The estimated transition dynamic can be expressed as $W^T\phi(x, a)$, where W^T is the last iterate.

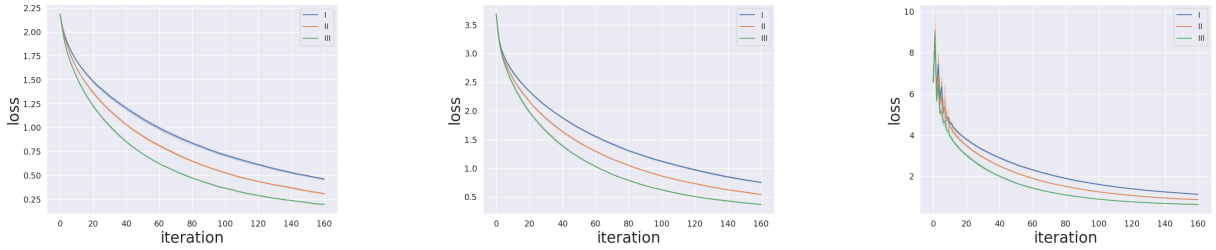
In Phase 2, we use the SPEDE algorithm (Ren et al., 2022), which is a planning algorithm based on the observation that under Gaussian noise, the linear spectral feature of the corresponding Markov transition operator can be obtained in a closed form. Moreover, SPEDE is suitable for a continuous state and action space. Note that in the original implementation, SPEDE samples transition functions from its posterior distribution at each episode, but, in our case, we do not need such a sampling.

We compare our method with a natural baseline: ordinary regression. For a fair comparison, we perform ordinary regression using the feature map $\phi(x, a)$. Let $J_k = \{(x_h^k, a_h^k, x_{h+1}^k)\}_h$ be the trajectory that includes samples in the k-th episode. The baseline estimator for the transition model, W^{baseline}, is defined as:

$$W^{\text{baseline}} := \underset{W}{\operatorname{argmin}} \Big\{ \sum_{k=1}^{K} \Big(\sum_{h=0}^{H} \|W\phi(x_h, a_h) - x_{h+1}\|_2^2 \Big) \Big\}.$$

Note that ordinary regression does not take the IV z_h and the UC e_h into consideration. We also use SPEDE as the planning component of the baseline algorithm.

We show the results in figures and tables. Table 1 indicates the instrument strength of different dynamics. Table 2 shows the computational time of different dynamics for estimation phase with a sample size of 5. We can observe that the average computation time for estimating dynamics under parametric setting is short. Figure 3 shows how the instrument strength affects the convergence rate of the gradient descent ascent in Phase 1. The loss here is equal to the difference between the true value W^* and the estimated value W^t after each iteration. It is evident that with stronger instrument strength the loss decreases faster than with weaker instrument strength. This phenomenon is consistent with our theory and intuition. With appropriately chosen instrument



(a) $k = 5$ (b) $k = 10$ (c) $k = 20$

Figure 3: Experiment results for the parametric setting. The gradient descent ascent loss $\|W^t - W^*\|_F$ for different settings of instrument strength under different dimensions

index \ dimensions	5	10	20
I	0.14	0.14	0.13
II	0.19	0.18	0.17
III	0.24	0.24	0.23
IV	0.14	0.14	0.13
V	0.18	0.18	0.18
VI	0.24	0.24	0.23

Table 1: Instrument Strength of different dynamics (parametric setting)

index \ dimension	5	10	20
I	4.06	4.53	4.90
II	3.90	4.47	4.86
III	3.71	4.44	4.89
IV	7.33	4.03	4.94
V	7.00	4.34	5.44
VI	6.23	3.94	4.70

Table 2: Average computation time (in seconds) for estimation phase (parametric setting)

index \ dimensions	5	10	20
Baseline	590.19 ± 29.14	832.92 ± 15.15	2257.49 ± 134.16
I	14.84 ± 1.31	173.98 ± 3.97	505.63 ± 8.86
II	4.17 ± 1.18	38.30 ± 3.04	290.65 ± 9.08
III	1.44 ± 1.30	29.52 ± 2.89	117.20 ± 6.48
IV	17.21 ± 1.43	100.79 ± 2.92	422.14 ± 18.41
V	8.00 ± 1.04	20.97 ± 3.72	290.41 ± 14.28
VI	6.36 ± 1.02	19.24 ± 3.71	140.58 ± 4.23

Table 3: Regret of different dynamics (parametric setting)

variables with sufficient instrument strength, we can still identify transition dynamics with small samples. Figure 4 plots the reward and its 95% confidence interval obtained by the SPEDE planning algorithm. The curve labeled 'opt' represents the policy derived from planning with the true underlying transition function. We observe that the reward of baseline decreases during training due to a wrong transition dynamic estimated through ordinary regression. This shows that in the presence of UCs, not only does ordinary regression produce biased estimates of the transition model, but also such estimation error will be propagated to planning and further amplified due to the sequential nature of the problem, producing a poor policy. Compared to baseline, our algorithm performs well in this case, and we summarize the results in Table 3 with a sample size of 5. We used the observations to compute uncentered covariance matrices, defined in (3.5), and instrument strength (IS), defined in (4.1). The regret is defined as the decrease in reward gained due to the execution of the policy produced by planning with the estimated transition function instead of the execution of the optimal policy produced by planning with the true transition function. In our experiment, we used the same online setting for different settings of IS, which allows us to compare the performance of the transition functions across different settings. We can observe that the regret for the baseline is large. It can also be seen that the influence of UCs is magnified when the dimensions increase. In low dimensions, the regret with weak instrument strength gets close to 0. In high dimensions, the regret gap between our algorithm and the optimal reward becomes larger because of not only the influence of UCs but also the larger reward at every step. However, it is obvious that the reward of all examples with sufficient instrument strength is still close to the optimal reward, which means the estimated transition function is a good estimator. In addition, our algorithm is better than the baseline. The robustness check experiment results is shown in Figure 5. We observe that the reward of all examples with sufficient instrument strength is still close to the optimal reward under the general covariance matrices setting. This consistency suggests that the estimated transition function is an effective estimator, and the proposed method is robust.

5.2 Nonparametric Case

We now consider examples with non-linear transition dynamics. In this case, we do not assume that the true transition function F^* lies in a known finite-dimensional space. We use our algorithm in 1-dimension, 3-dimension, and 5-dimension. In order to compare results in different dimensions, we use similar parameter setting in different dimensions. We let k stands for the number of dimensions. Specifically, we consider a CMDP-IV that operates in the following spaces: $\mathcal{S} = \mathcal{U} = \mathcal{Z} = \mathbb{R}^k$, $\mathcal{A} = [-1, 1]^k$. We generate 80 episodes and each episode has a horizon $H = 500$. The environment starts from an zero vector $x_0 = \mathbf{0}$, generates a sequence of UCs $\{e_h\}_h \overset{\text{iid}}{\sim} \mathcal{P}_e$ and a sequence of observable IVs $\{z_h\}_h \overset{\text{iid}}{\sim} \mathcal{P}_z$. As in the parametric setting, e_h and z_h are generated from different Gaussian distributions: $\mathcal{P}_e = \mathcal{N}(\mu_e, \Sigma_e)$, $\mathcal{P}_z = \mathcal{N}(\mu_z, \Sigma_z)$. We use a non-linear transition function $F^*(x_h, a_h) = \ln(|x_h - 1| + 1) - \frac{a_h}{2}$. At the h-th step, given the current state x_h, the behavior policy samples the action a_h, and then the next state x_{h+1} is generated by the equations

$$a_h \sim \text{Proj}_{[-1,1]^k}\left(\mathcal{N}(z_h + e_h, \Sigma_a)\right), \quad x_{h+1} = F^*(x_h, a_h) + e_h = \ln(|x_h - 1| + 1) - \frac{a_h}{2} + e_h ,$$

where Σ_a is a diagonal matrix with all diagonal elements equal to one half. The action a_h is generated by first sampling from a Gaussian distribution and then projecting the sample onto the interval $[-1, 1]^k$. The projection is used to stabilize the state dynamics of the behavior policy. The reward function is $r_h(x_h, a_h) = -0.05\|x_h\|^2$. The optimal policy forces the agent to get close to zero vector, since this state has the highest reward. We use different variance Σ_e, Σ_z to control the instrument strength and study its influences on our algorithm. In our experiment, we fix Σ_e equals

index \ dimensions	1	3	5
I	0.0024	0.0015	0.0011
II	0.0044	0.0027	0.0019
III	0.0064	0.0038	0.0028
IV	0.0012	0.0007	0.0005
V	0.0026	0.0015	0.0011
VI	0.0041	0.0024	0.0017

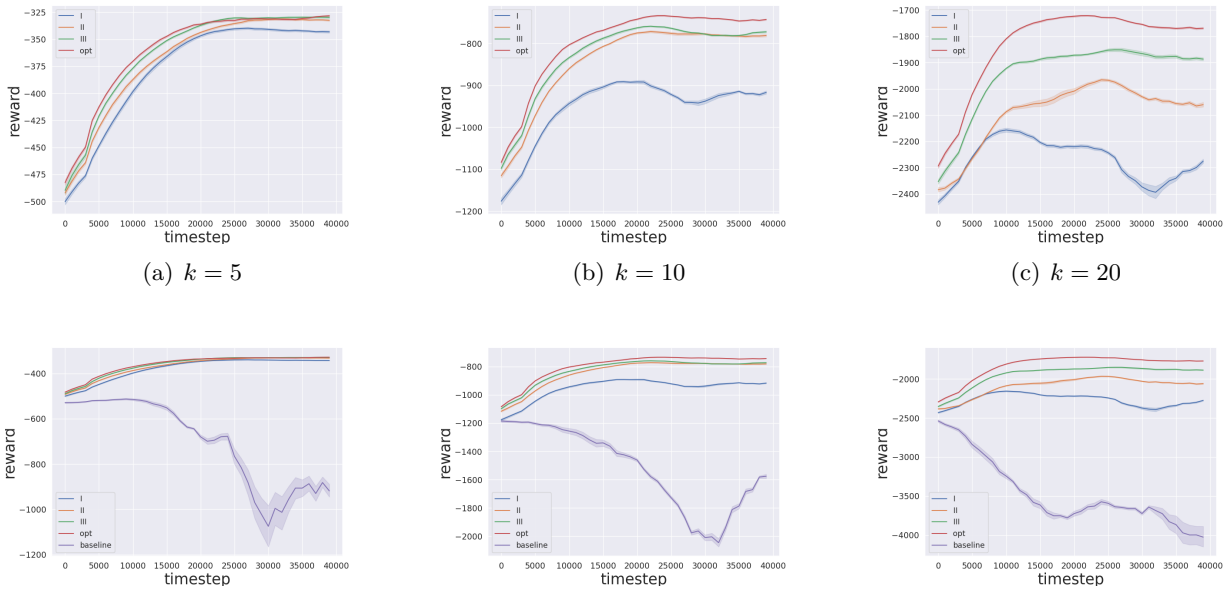
Table 4: Instrument Strength of different dynamics (nonparametric setting)

index \ dimension	1	3	5
I	5.56	4.75	4.05
II	5.38	5.08	3.15
III	5.03	4.81	3.43
IV	7.23	8.05	5.45
V	6.68	8.43	5.78
VI	7.12	7.21	6.23

Table 5: Average computation time (in seconds) for estimation phase (nonparametric setting)

index \ dimensions	1	3	5
Baseline	11.92 ± 0.91	47.85 ± 0.77	74.74 ± 1.01
I	2.54 ± 0.27	16.04 ± 0.55	35.90 ± 0.87
II	2.35 ± 0.35	11.93 ± 0.82	27.14 ± 0.77
III	1.73 ± 0.29	6.83 ± 0.56	19.98 ± 0.99
IV	3.22 ± 0.19	13.25 ± 0.40	27.29 ± 0.55
V	2.32 ± 0.16	13.20 ± 0.46	26.72 ± 0.53
VI	1.89 ± 0.22	9.02 ± 0.41	24.15 ± 0.61

Table 6: Regret of different dynamics (nonparametric setting)



(a) $k = 5$ (b) $k = 10$ (c) $k = 20$

(d) $k = 5$ (e) $k = 10$ (f) $k = 20$

Figure 4: Experiment results for the parametric setting. Top panel: The performance curves (with 95% confidence interval) of reward versus the time steps for different transition functions (without baseline). Bottom panel: The performance curves (with 95% confidence interval) of reward versus the time steps for different transition functions (including the ordinary regression baseline). The time step is the episode for SPEDE.

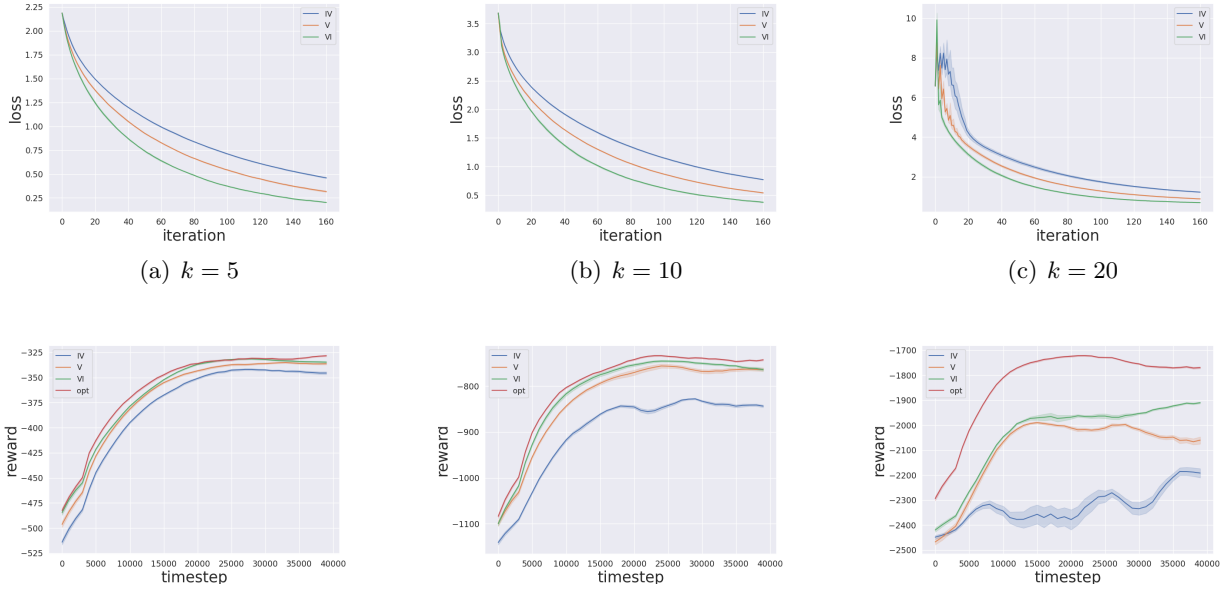
to identity matrix and use the following variances for z_h:

$$\text{I) } \Sigma_z = 0.5 I_k; \quad \text{II) } \Sigma_z = 0.9 I_k; \quad \text{III) } \Sigma_z = 1.5 I_k.$$

Similar to the parametric setting, we introduce experiments which incorporate variations in the covariance matrices Σ_a and Σ_e. In the experiment, we use the following pairwise covariance matrices:

$$\begin{aligned} \text{IV) } & \Sigma_z = 0.5 I_k, \Sigma_e = T(-0.5) \text{ and } \Sigma_a = T(-0.3); \\ \text{V) } & \Sigma_z = 0.9 I_k, \Sigma_e = T(0.1) \text{ and } \Sigma_a = T(0); \\ \text{VI) } & \Sigma_z = 1.5 I_k, \Sigma_e = T(0.5) \text{ and } \Sigma_a = T(0.3). \end{aligned}$$

We instantiate Algorithm 1 with polynomial feature maps $\phi(x, a) = [1, x, a, x^2, a^2]^\top$ and $\psi(x, z) = [1, x, z, x^2, z^2]^\top$. Note that our transition function is a logarithmic function with some noise, so it does not lie in the finite-dimensional space spanned by the chosen feature maps. We use the minibatch stochastic gradient descent to compute W^{sad}. We use zero matrices to initialize estimates K_0 and W_0. At t-th iteration, stepsizes we use are $\eta_t^\theta = \frac{1}{16+t}$ and $\eta_t^\omega = \frac{1}{16+t}$. The estimated transition function can be expressed as $W^T \phi(x, a)$. In Phase 2, we use the SPEDE planning algorithm. We compare our method with ordinary regression. For a fair comparison, we do ordinary regression on the feature map $\phi(x, a)$. We use the same baseline estimator as specified in the parametric setting. Note that ordinary regression does not take into account the IV z_h and the UC e_h. We also use the SPEDE algorithm for planning.

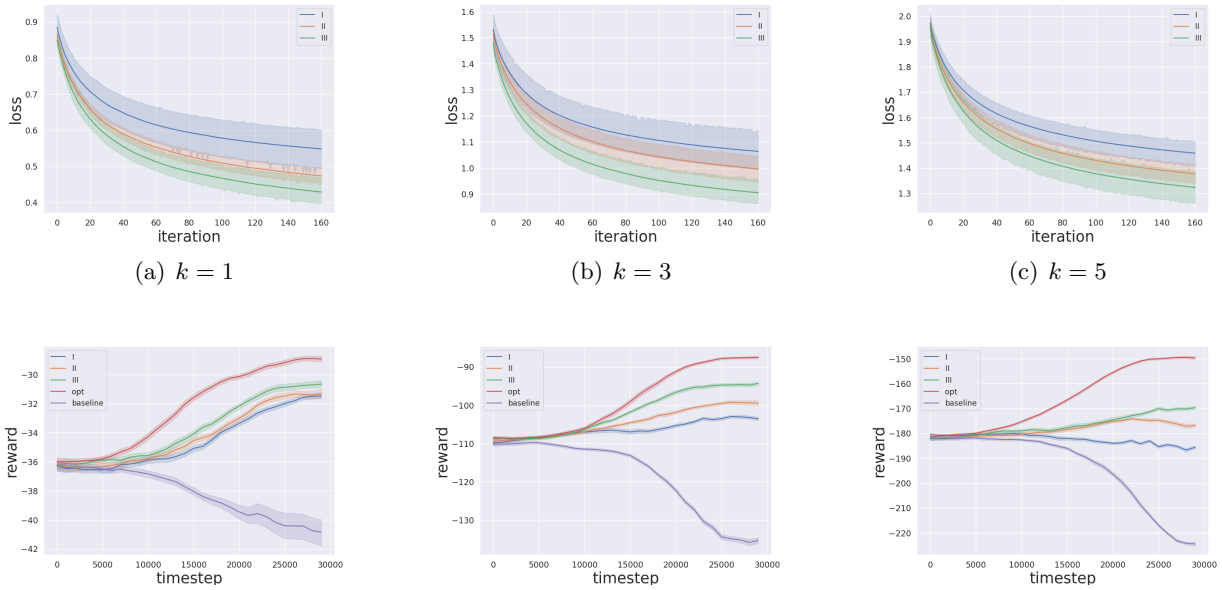


(a) $k = 5$ (b) $k = 10$ (c) $k = 20$

(d) $k = 5$ (e) $k = 10$ (f) $k = 20$

Figure 5: Robustness check experiment results for the parametric setting. Top panel: The gradient descent ascent loss $\|W^t - W^{sad}\|_F$ for different settings of instrument strength under different dimensions. Bottom panel: The performance curves (with 95% confidence interval) of reward versus the time steps for different transition functions (without baseline). The time step is the episode for SPEDE.

Table 4 shows the instrument strength of different dynamics. Table 5 shows the computational time for estimation phase. Figure 6 shows the result in the nonparametric setting. The top panel shows how the strength of the instrument affects the GDA convergence rate in Phase 1. We monitor the progress of GDA by plotting the difference between W^{sad}, which is obtained in closed form in (3.7), and estimated W^t after each iteration. We observe that the loss decreases faster with stronger instruments. The bottom panel shows the reward and its 95% confidence interval obtained by the SPEDE planning algorithm. Similarly, the curve marked as 'opt' corresponds to the policy generated by planning with the actual underlying model. The reward of the baseline does not converge to the optimal reward. In the presence of UCs, the bias in the ordinary regression produces biased estimates whose error is propagated to planning and amplified due to the sequential nature of the problem, resulting in a low-performing policy. Table 6 summarizes the results for a sample size of 5. We calculate instrument strength and regret in the same way as in the parametric setting, by computing the smallest nonzero eigenvalue of related covariance matrices. Compared to the baseline, our algorithm has a good performance that improves with the strength of the instrument. There is a regret gap between our algorithm and the optimal regret, as the exact transition function can not be found in a finite-dimensional space. This is predicted by Theorem 14. We also observe that the regret gap between our algorithm and the optimal reward becomes larger as the dimension of space increases. This phenomenon is due to not only the larger reward at each step in higher dimensions but also the amplified influence of UCs, which lower the quality of the estimated function. However, it is obvious that the reward of all examples with sufficient instrument strength is still close to the optimal reward, which means the estimated transition function is a good estimator. We present



(a) $k = 1$ (b) $k = 3$ (c) $k = 5$

(d) $k = 1$ (e) $k = 3$ (f) $k = 5$

Figure 6: Experiment results for the nonparametric setting. Top panel: The gradient descent ascent loss $\|W^t - W^{sad}\|_F$ for different settings of instrument strength under different dimensions. Bottom panel: The performance curves (with 95% confidence interval) of reward versus the time steps for different transition functions. The time step is the episode for SPEDE.

the results of the robustness check experiments in Figure 7. We observe that our algorithm is still better than the baseline, which indicates the robustness of the method.

5.3 Assessment with MovieLens Dataset

We construct a semi-synthetic data based on MovieLens 1M dataset (Harper and Konstan, 2015), MovieLens is a dataset that records people's ratings for different movies and it contains approximately 1 million ratings (on the scale 0 – 5) of 3952 movies created by 6040 individuals. The rating matrix is a highly sparse matrix, containing very few user/movie rating pairs.

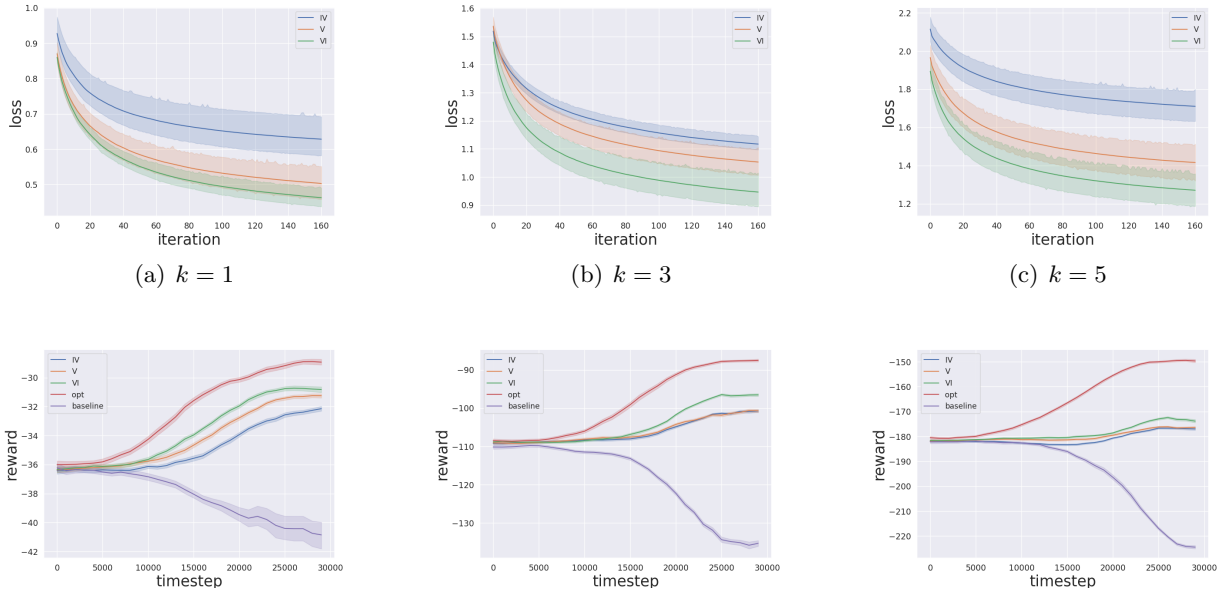
We now describe our semi-synthetic setup based on the user/movie rating pairs. Let R denote the rating matrix and let $R = \widehat{U}\Sigma\widehat{V}^\top$ the SVD of the rating matrix R. The rows of the matrix $\widehat{U}$ represent the preference of each user for different movie categories, and the rows of the matrix $\widehat{V}$ represent the membership of a movie in these categories. We focus on the top 10 categories and 10 movies selected by singular value decomposition (SVD). We keep the top 10 singular values in matrix Σ and the corresponding leading singular vectors in $\widehat{U}$ and $\widehat{V}$. We denote the resulting matrices by U and V, respectively. Finally, we use $\widetilde{V}$ to denote the first 10 rows of V. The matrix $\widetilde{V}$ is a 10 by 10 matrix of movies by categories.

Figure 8 shows the causal diagram that represents data generation in our semi-synthetic application. To construct a Markov decision process, we let the state be the user's preference (captured by ratings for movie categories), and the action be whether the user watches these movies. In this application, the IV is the recommendation from the recommender system. In recommender systems, the recommendation is sufficiently randomized conditionally on the user's characteristics and only affects preference by encouraging the user to watch the recommended movie. The UC in this setting are the unobserved factors that affect both whether the user watches the recommended movie (the action) and his preference (the state variable) simultaneously. For example, the director of a movie could affect the action (how likely the user actually watches the movie) and the state variable (the updated preference of movie categories after watching) at the same time. Our goal is to identify a sequence of movie recommendations that, when actually followed by users, will result in high ratings for movies.

Formally, at the beginning of an episode, we randomly choose one user's initial ratings for the 10 movie categories from the rows of matrix U (with Gaussian noise) as the initial state variable $x_0 \in \mathbb{R}^n$. We let a sequence of UCs $\{e_h\}_h$ follow a Gaussian distribution, that is, $\{e_h\}_h \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{10})$, a sequence of i.i.d. IVs, $z_h \in \mathbb{R}^{10}$, $h = 1, \ldots, H$ follows a multinomial distribution with n total recommendations, and probability of recommending the i-th movie p_i. Note that the IV is independent of the current state and UC. At the h-th step, the behavior policy samples the action a_h, and then the next state x_{h+1} is generated as

$$a_{h,i} \sim \text{Bernoulli}\left(\frac{1}{1 + e^{-2z_{h,i} - 0.1x_{h,i} + 0.8e_{h,i}}}\right), \quad x_{h+1} = F^*(x_h, a_h) + e_h = x_h + \widetilde{V}^\top a_h + e_h \,.$$

The action $a_{h,i}$, the i-th entry of a_h, is generated by sampling from a Bernoulli distribution, where the parameter is a logistic transform of a linear combination of instruments, states and confounders. The transition function mimics how a user would update her preference after watching a movie. If a user watches one particular movie, we add movie's category to the current user's preference, and if the user does not watch any movies, we do not update the preference. The reward function is $r_h(x_h, a_h) = \|\text{Proj}_{[0,5]}\left(\Sigma\widetilde{V}^\top x_h\right)\|_1$, representing the sum of ratings. We generate 200 episodes, each with horizon $H = 100$. We use different multinomial distributions to control the strength of the instrument and study its influences on our algorithm. Under the above setup, the sequences $\{z_h\}_h$



(a) $k = 1$ (b) $k = 3$ (c) $k = 5$

(d) $k = 1$ (e) $k = 3$ (f) $k = 5$

Figure 7: Robustness check experiment results for the nonparametric setting. Top panel: The gradient descent ascent loss $\|W^t - W^{sad}\|_F$ for different settings of instrument strength under different dimensions. Bottom panel: The performance curves (with 95% confidence interval) of reward versus the time steps for different transition functions. The time step is the episode for SPEDE.

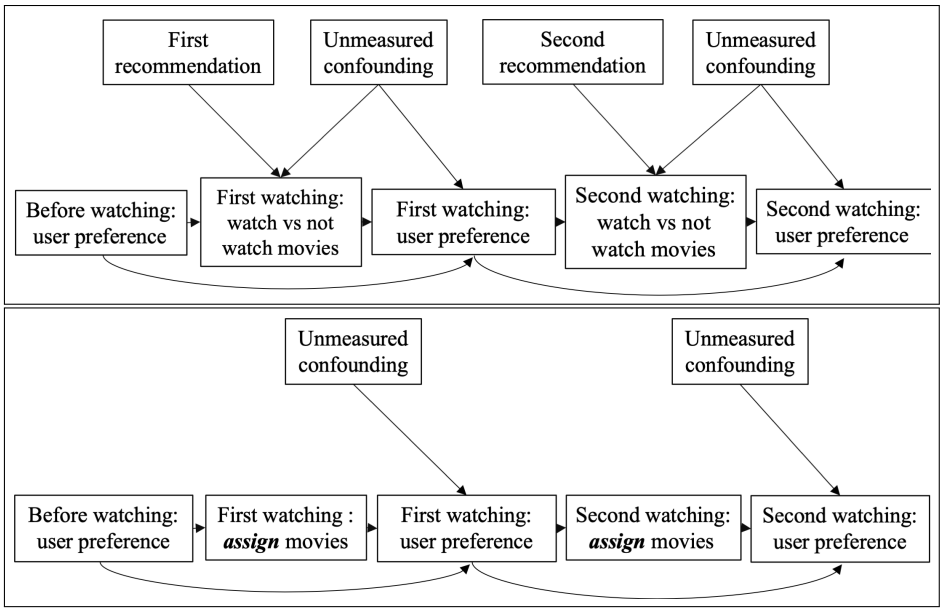


Figure 8: The recommender application. Top panel: DAG representing data generation process in a recommender system where UCs are present. Bottom panel: DAG representing a recommender system in action.

and $\{e_h\}_h$ are both i.i.d. and independent of the current state variable. Therefore, z_h is a valid IV and e_h is a valid UC in the proposed dynamics. We use three different multinomial distributions:

I. $n = 10$, $p_1 = p_2 = \cdots = p_4 = \frac{1}{5}$, $p_5 = p_6 = \cdots = p_{10} = \frac{1}{30}$;

II. $n = 10$, $p_1 = p_2 = \frac{1}{5}$, $p_3 = p_4 = \cdots = p_6 = \frac{1}{10}$, $p_7 = p_8 = \cdots = p_{10} = \frac{1}{20}$;

III. $n = 10$, $p_1 = p_2 = \cdots = p_{10} = \frac{1}{10}$.

We also consider a revised data generation process based on the fact that recommendation is usually made based on user preference. The difference from the model above is the extra arrow from user preference to recommendation. In the recommender system context, this extra arrow represents that the recommendation is made based on user preference. The sequence of IVs, $z_h \in \mathbb{R}^{10}$, $h = 1, \ldots, H$ still follows a multinomial distribution with n total recommendations, but the probability of recommending the i-th movie p_i is determined by the estimated ranking of the i-th movie based on the current user preference. To be specific, we have a multinomial distribution with probability $q_1, \cdots, q_{10}$, where $q_1 \geq q_2 \geq \cdots \geq q_{10}$. At the h-th step, the ranking of the i-th movie can be estimated as the i-th entry of $\mathrm{Proj}_{[0,5]}\left(\Sigma \widetilde{V}^\top x_h\right)$. We sort the estimated ranking in decreasing order and suppose the ranking of the i-th movie is the j-th highest ranking. Then, we let $p_i = q_j$. We use three different multinomial distributions:

I. $n = 10$, $q_1 = q_2 = q_3 = \frac{1}{3}$, $q_4 = q_5 = \cdots = q_{10} = 0$;

II. $n = 10$, $q_1 = q_2 = \frac{1}{5}$, $q_3 = q_4 = \cdots = q_6 = \frac{1}{10}$, $q_7 = q_8 = \cdots = q_{10} = \frac{1}{20}$;

III. $n = 10$, $q_1 = q_2 = \cdots = q_{10} = \frac{1}{10}$.

In the three offline dynamics only the distribution of IV is different; their corresponding evaluation dynamics are the identical. This enables us to evaluate how the strength of instrument affects estimation of optimal policy in the evaluation dynamics.

We instantiate Algorithm 1 with feature maps $\phi(x,a) = [1,x,a]^\top$ and $\psi(x,z) = [1,x,z]^\top$. Note that the current state x is also a variable in the feature map ψ. With this feature map, the true transition function can be written as $W^*\phi(x,a)$, where $W^* = [I, \widetilde{V}^\top]$. We use the minibatch stochastic gradient descent to compute W^{sad}. We set initial estimates K_0 as a zero matrix and W_0 as a matrix where every entry is one fifth. At t-th iteration, stepsizes we use are $\eta_t^\theta = \frac{1}{550+t}$ and $\eta_t^\omega = \frac{1}{1800+t}$. For the revised recommender application, stepsizes are $\eta_t^\theta = \frac{1}{600+t}$ and $\eta_t^\omega = \frac{1}{1800+t}$. The estimated transition function can be expressed as $W^T\phi(x,a)$. In Phase 2, we use the SPEDE planning algorithm.

We compare our method with ordinary regression where we use the feature map $\phi(x,a)$ for a fair comparison. We use the same baseline estimator as specified in the parametric setting. Recall that ordinary regression does not take IV z_h and UC e_h into consideration. We use the SPEDE planning algorithm to recover the optimal policy.

Table 7 shows the instrument strength of different dynamics. Table 8 shows the computational time for the estimation phase. Figure 9 shows the results on the MovieLens 1M dataset. The top panel shows how the strength of the instrument affects the convergence of GDA in Phase 1. The estimation loss, measured by $\|W^t - W^*\|_F$, decreases faster as the strength of the instrument increases. We also plot the loss for the baseline in the top panel. We note that our method achieves a smaller loss when estimating the transition function compared to the baseline. The bottom panel plots the reward and its 95% confidence interval obtained by the SPEDE planning algorithm. We observe that the baseline reward is lower than the optimal reward. This shows that in the presence of UC, a poor policy is produced due to the bias in the ordinary regression estimates, as well as the increased estimation error. We summarize the results in Table 9. We compute instrument strength by averaging the five smallest nonzero eigenvalues from uncentered covariance matrices of the feature maps ϕ and ψ. The regret is calculated in the same way as in the parametric setting. Compared to the baseline, the policies obtained by the IV-based estimation all achieve higher rewards than the one obtained by OLS. Moreover, these policies have a similar reward as the policy obtained by planning with the true underlying model (the curve labeled 'opt'). We also observe that there is a very minor numerical difference between the original model and the revised model.

6. Conclusion and Discussions

Our model is motivated by real-world applications of RL in recommender systems and healthcare, where UCs are present. We show that, for additive nonlinear transition dynamics, a valid IV can help identify the confounded transition function. The proposed IVVI algorithm is based on a primal-dual formulation of the conditional moment restriction implied by the IV. Moreover, our stochastic approximation approach to the nonparametric IV problem is of independent interest. We derive the convergence rate of IVVI. Furthermore, we derive the sample complexity of offline RL with IVs in the presence of unmeasured confounders.

An Updated Survey

After this paper has been posted publicly, there is an emerging literature on the use of causal inference in RL. Thanks to the suggestion by an anonymous reviewer, we present an updated survey for these works, even though some of these works come after this paper.

index \ model	Original Model	Revised Model
I	0.012	0.002
II	0.018	0.019
III	0.020	0.021

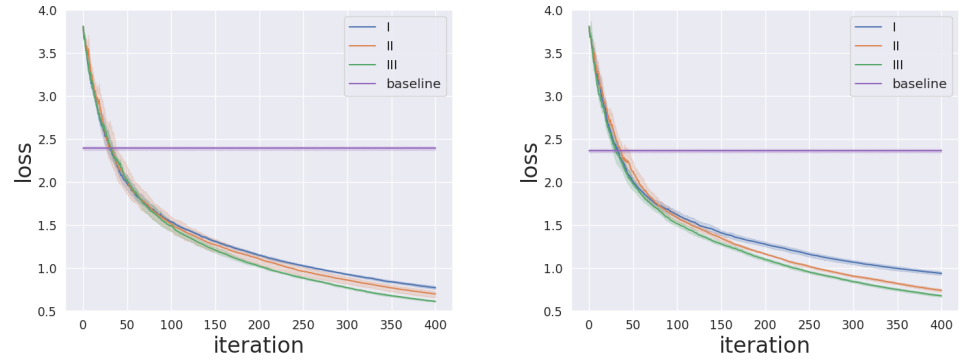
Table 7: Instrument Strength of different dynamics (MovieLens dataset)

index \ model	Original Model	Revised Model
I	4.92	3.18
II	4.21	3.33
III	4.25	3.13

Table 8: Average computation time (in seconds) for estimation phase (MovieLens dataset)

index \ model	Original Model	Revised Model
Baseline	149.98 ± 79.84	137.47 ± 76.56
I	76.24 ± 81.27	110.88 ± 73.90
II	46.20 ± 75.71	89.43 ± 74.68
III	18.30 ± 75.32	85.34 ± 81.11

Table 9: Regret of different dynamics (MovieLens dataset)



(a) Original Model (b) Revised Model

(c) Original Model (d) Revised Model

Figure 9: Experiment results on the MovieLens dataset. Top panel: The gradient descent ascent loss $\|W^t - W^{sad}\|_F$ for different settings of instrument strength under different data generation process. Bottom panel: The performance curves (with 95% confidence interval) of reward versus the time steps for different transition functions. The time step is the episode for SPEDE.

The work by Wang and Tchetgen Tchetgen (2018) considers average treatment estimation under unobserved confounders. They study binary IVs and binary treatment, derive conditions for ATE identification and propose a multiple robust estimator based on existing regression based, inverse probability weighting based, and G-formula based estimators. Building upon the paper by Wang and Tchetgen Tchetgen (2018), Cui and Tchetgen Tchetgen (2021) consider estimation of optimal treatment in the presence of unmeasured confounders. They look at settings with binary IVs and binary treatments. With additional assumptions on the IVs (the no unmeasured common effect modifier or the independent compliance type conditions in their paper), the authors derive identification results for the optimal treatment and proposed multiply robust classification-based estimators. Different from our work, the model in the above two papers is one-stage and thus not sequential model. Our stochastic apporoixmation-based estimator is computationally attractive. The very recent work by Bilodeau et al. (2022) studies the problem of online learning an low-regret dynamic policy. They consider a bandit setting where the reward depends on the action only. Interestingly, they allow the presence of unobserved confounder that affects both the action and the

reward. Different from our paper, we consider an MDP setting where reward depends on the action and the state variable. They consider discrete action space while we work with continuous action space."

There is a line of work that explore the use of IV in RL. Chen and Zhang (2023) use existing partial identification results for IV and study policy improvement under the binary-treatment binary-instrument setting. Different from our work, our paper studies the continuous-treatment continuous-instrument case, and the transition function is identified by the condition moment restriction. Fu et al. (2022) focus on the discrete-treatment discrete-instrument case, show that by using certain weighting schemes the values of policies are identified, and adapt pessimistic offline RL algorithm to learn the optimal policy. The work of Xu et al. (2023) also study the discrete-treatment discrete-instrument case but, different from Fu et al. (2022), they derive the efficient influence function and propose a more efficient estimator for policy values. Different from these two works, our identification and estimation strategy is based on conditional moment restrictions. Yu et al. (2022) study a novel learning setting which they termed strategic MDP. Their model features the strategic interactions between a principal and a sequence of myopic agents with private types. They show that IV structure exists in the model. While we work under different learning settings, both their work and ours use conditional moment restrictions for identification and estimation.

Researchers have also explored other causal structure in RL. For example, Wang et al. (2021) studies confounded MDP where front-door or back-door adjustments are available, Bennett and Kallus (2023) studies the case where proxy variables are present. Even in the IV case, our work has inspired several related work to explore the use of IV for other tasks in RL, such as offline policy evaluation Xu et al. (2023), offline policy learning in strategic MDP Yu et al. (2022), policy improvement Chen and Zhang (2023) and offline RL with discrete instruments Fu et al. (2022). What we aim to show is the CMDP-IV model we propose is a reasonable model for quite a few datasets, and the estimation method is backed by both theoretical proofs and synthetic and semi-synthetic experiments.

Relaxation of Assumption A.1

There are several directions to proceed from Assumption A.1.

For example, we allow instruments to depend on previous history. Concretely, we allow z_h to depend on $\{x_h, \ldots, x_1, z_{h-1}, \ldots, z_1\}$. Consider

$$
\begin{aligned}
&x_{h+1} = F(x_h, a_h) + e_h,\ a_h \sim \pi_h(\cdot|x_h, z_h, e_h),\\
&\{e_1, \ldots, e_H, x_1\} \text{ are independent}, x_1 \sim \xi,\ e_h \sim N(0,1) \text{ for all } h\\
&z_1 \sim \mathcal{P}_{z,1}(\cdot|x_1),\ z_h \sim \mathcal{P}_{z,h}(\cdot|x_h, \ldots, x_1, z_{h-1}, \ldots, z_1) \text{ for all } h.
\end{aligned}
$$

By the same reasoning as above, a conditional moment restriction will be implied and estimation can be done. Let L_h be the law of (x_h, a_h, z_h, x_{h+1}), and $L(x, a, z, x')$ be the average mixture of $\{L_1, \ldots, L_H\}$. Let $p_{x,z,h}(x, z)$ be the marginal of (x_h, z_h). Then the density of L_h is

$p_{x,z,h}(x,z)\mathcal{P}_e(e)\pi_h(a|x,z,e)1(z'=F^*(x,a)+e)$, and the density of L is

$$\begin{aligned}
&\frac{1}{H}\sum_{h=1}^{H} p_{x,z,h}(x,z)\mathcal{P}_e(e)\pi_h(a|x,z,e)1(z'=F^*(x,a)+e)\\
&=\mathcal{P}_e(e)\underbrace{\left(\frac{1}{H}\sum_{h=1}^{H} p_{x,z,h}(x,z)\right)}_{=:p_{x,z}(x,z)}\underbrace{\left(\sum_{h=1}^{H}\frac{p_{x,z,h}(x,z)}{\sum_{h'=1}^{H}p_{x,z,h'}(x,z)}\pi_h(a|x,z,e)\right)}_{=:\bar{\pi}(a|x,z,e)}\\
&\quad\cdot 1(z'=F^*(x,a)+e)
\end{aligned}$$

Clearly for $(x,a,z,e,x')\sim L$ we have $\mathbb{E}[F^*(x,a)-x'|z]=0$.

We do not aim to exhaust all possibilities in the paper because we mainly aim to develop the idea that in the confounded MDP setting, one could identify and estimate transition dynamics and thus the optimal policy through conditional moments implied by instrumental variables.

Episode-wise dependence is also possible. Let $\mathcal{F}_t=\sigma\{x_{h,\tau},z_{h,\tau},a_{h,\tau},x'_{h,\tau}\}_{h\in[H],\tau=1,\dots,t}$ be the data of the first t episodes. Then we could let $x_{1,t+1}\sim\xi(\cdot|\mathcal{F}_t)$, i.e., the first state at the $(t+1)$-th episode can be chosen depending on previous t episodes. This is particularly relevant when the behavior policy is updated in the observation process.

The additive structure seems hard to relax in our opinion. A possible extension is as follows. For ease of notation let the state space $\mathcal{X}$ be $\mathbb{R}$. Then for a transition function $F:\mathbb{R}\times\mathcal{A}\times\mathbb{R}\to\mathbb{R}$, the observation dynamics writes

$$\begin{aligned}
&x_{h+1}=F(x_h,a_h,e_h),\ a_h\sim\pi_h(\cdot|x_h,z_h,e_h),\\
&\{e_1,\dots,e_H,x_1,z_1,\dots,z_H\}\text{ are independent,}\\
&x_1\sim\xi,\ e_h\sim N(0,1),z_h\sim\mathcal{P}_z\text{ for all }h.
\end{aligned}$$

Following Chernozhukov et al. (2007), assume $F(x,a,\cdot)$ is strictly increasing on $\mathbb{R}$ for all (x,a). Let L_h be the law of (x_h,a_h,z_h,x_{h+1}), and $L(x,a,z,x')$ be the average mixture of $\{L_1,\dots,L_H\}$. Then by the same argument as that of Chernozhukov et al. (2007), for any $\tau\in(0,1)$, it holds

$$\mathbb{E}\big[1\big(x'<F(x,a,\tau)\big)-\tau\big|z\big]=0 \tag{6.1}$$

where the expectation is taken w.r.t $(x,a,z,x')\sim L$. However, having arrived at a conditional moment restriction, we notice several difficulties to proceed. First, monotonicity is imposed on the transition function $F(x,a,\cdot)$, and it is unclear how to test this assumption on real data or exploit it for estimation. Second, the indicator function present in the conditional moment restriction is nonsmooth, bringing challenges to theoretical analysis. Traditional estimation methods in existing literature are not suitable for RL applications. Consider the special case of single decision stage ($H=1$), which is essentially the nonparametric quantile IV (NPQIV) problem in the econometric literature. There, most papers use traditional computation-heavy nonparametric estimators such as sieve or kernel estimators (Chernozhukov et al., 2007; Chernozhukov and Hansen, 2005; Horowitz and Lee, 2007), which are not suitable for RL applications where online procedures are definitely preferred.

Comment on Assumption A.2

The i.i.d. data assumption is purely for simplifying notations. We just replace anywhere in Algorithm 1 where we evaluate a function $f(x,z,a,x')$ at a point (x,z,a,x') drawn from the average visitation

distribution $\bar{d}_{\pi_b}$ with $\frac{1}{H}\sum_{h=1}^{H} f(x_h, z_h, a_h, x_{h+1})$, where $\{x_h, z_h, a_h, x_{h+1}\}_h$ follows the dynamics describe by $x_{h+1} = F^*(x_h, a_h) + e_h$ and $a_h \sim \pi_h(\cdot|x_h, z_h, e_h)$.

Concretely, Lines 3-4 in Algorithm 1

$$\phi_t \leftarrow \phi\left(x_t, a_t\right), \psi_t \leftarrow \psi\left(z_t\right)$$
$$W_{t+1} \leftarrow W_t - \eta_t^\theta \cdot \left(K_t \psi_t \phi_t^\top\right), \quad K_{t+1} \leftarrow K_t + \eta_t^\omega \cdot \left(K_t \psi_t \psi_t^\top + x_t' \psi_t^\top - W_t \phi_t \psi_t^\top\right).$$

should be replaced with

$$W_{t+1} \leftarrow W_t - \eta_t^\theta \cdot \left(K_t \left(\frac{1}{H}\sum_{h=1}^{H} \psi(z_{h,t})\phi(x_{h,t}, a_{h,t})^\top\right)\right) \tag{6.2}$$
$$K_{t+1} \leftarrow K_t + \eta_t^\omega \cdot \left(\frac{1}{H}\sum_{h=1}^{H}\left(K_t \psi(z_{h,t})^{\otimes 2} + x_{h+1,t}\psi(z_{h,t})^\top - W_t \phi(x_{h,t}, a_{h,t})\psi(z_{h,t})^\top\right)\right).$$

where $(x_{h,t}, z_{h,t}, a_{h,t}, x_{h+1,t})$ are the h-th step data in the t-th episode.

The proof of convergence for the stochastic approximation procedure goes through by the following reasoning. Suppose we are at the t-th iteration of the primal-dual algorithm, then conditional on the previous $(t-1)$ iteration (i.e., data from the previous $t-1$ episodes), the gradients in Eq 6.2 are unbiased estimates of the corresponding population gradient.

Finally, we remark that in practice, such an i.i.d. sampling oracle from $\bar{d}_{\pi_b}$ can be approximated by the following sampler: draw k uniformly from $[K]$ and h uniformly from $[H]$, and then output $\{x_{h,t}, z_{h,t}, a_{h,t}, x'_{h,t}\}$ which is data at the h-th timestep in the t-th episode.

Acknowledgments

We sincerely thank the editor, Professor Eric Laber, and three anonymous reviewers for their constructive comments. We also thank the attendees of the TTIC Chicago Summer Workshop: New Models in Online Decision Making for Real-World Applications (2022) for many helpful discussions. YW was supported in part by the Office of Naval Research under grant number N00014-23-1-2590, and the National Science Foundation under Grant No. 2231174, and No. 2310831, No. 2428059, and a Michigan Institute for Data Science Propelling Original Data Science (PODS) grant. MK was supported in part by the National Science Foundation under Grant No. ECCS-2216912.

Appendix A. Proof Sketch

The proof consists of two parts: the analysis of the convergence of the stochastic gradient descent-ascent (Line 3–6) and the analysis of the planning phase using the estimated model (Line 8–11).

In Remark 10 we emphasized the stochastic minimax optimization problem is only strongly concave in the dual variable. This motivates us to study the recursion of the following asymmetric potential function. For some $\lambda > 0$, define

$$\widetilde{P}_t = \mathbb{E}\big[\|W_t - W^*\|_F^2\big] + \lambda \mathbb{E}\big[\|K_t - K^*(W_t)\|_F^2\big]$$

where $K^*(W) = (WA^\top - C)B^{-1}$ with A, B and C defined in Eq. (3.5). The matrix $K^*(W)$ is the optimal dual variable in the saddle-point problem Eq. (3.7) when the primal variable is fixed at W. In order to get around the assumption of bounded variance of stochastic gradients, which is common in the optimization literature Nemirovski et al. (2009), we follow the idea in the work of Nguyen et al. (2018) where we upper bounds the variance of stochastic gradients by the suboptimality of the current iterate; see Lemma 18. Thus our algorithm does not require projection in each iteration. A careful analysis of the recursion for the sequence $\{\widetilde{P}_t\}$ shows the error in squared Frobenius norm converges at the rate $O(1/t)$.

The second element in our analysis is the decomposition of difference of value functions, which is adapted from Lemma 4.2 of Cai et al. (2020).

Lemma 15 (Suboptimality Decomposition)**.** *It holds that for all states $x \in \mathcal{S}$,*

$$\begin{aligned} V_1^*(x) - V_1^{\widehat{\pi}}(x) = &\sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\iota_h(x_h, a_h) \,|\, x_1 = x] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\xi_h(x_h) \,|\, x_1 = x] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\iota_h(x_h, a_h) \,|\, x_1 = x], \end{aligned} \tag{A.1}$$

where $\widehat{\pi}$ is the output of Algorithm 1, the expectations $\mathbb{E}_{\pi^*}$ and $\mathbb{E}_{\widehat{\pi}}$ are taken over trajectories generated by policies π^* and $\widehat{\pi}$ under the true transition function F^*, respectively, $\xi_h = \langle \widehat{Q}_h, \pi_h^* - \widehat{\pi}_h \rangle_{\mathcal{A}}$ for all $x \in \mathcal{S}$, and $\iota_h = (r_h + \mathbb{P}\widehat{V}_{h+1}) - \widehat{Q}_h$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$.

Proof See Appendix C.3 for a detailed proof. ■

Appendix B. Structural Causal Model and Intervention

Structural Causal Models (SCMs) provide a formalism to discuss the concept of causal effects and intervention. We briefly review its definition in this section and refer readers to Pearl (2009, Ch. 7) for a detailed survey of SCMs.

A structural causal model is a tuple (A, B, F, P), where A is the set of exogenous (unobserved) variables, B is the set of endogenous (observed) variables, F is the set of structural functions capturing the causal relations, and P is the joint distribution of exogenous variables. An SCM is associated with a causal directed acyclic graph, where the nodes represent the endogenous variables and the edges represent the functional relationships. In particular, each exogenous variable $X_j \in B$ is generated through $X_j = f_j(X_{\mathrm{pa}_D(j)}, U_j)$ for some $f_j \in F$, $U_j \in B$, where $\mathrm{pa}_D(j)$ denotes the set of parents of X_j in D. A distribution over the endogenous variables is thus entailed.

An intervention on a set of endogenous variables $X \subseteq B$ assigns a value x to X while keeping untouched other exogenous and endogenous variables and the structural functions, thus generating

a new distribution over the endogenous variables. We denote by $\mathrm{do}(X = x)$ the intervention on X and write $\mathrm{do}(x)$ if it is clear from the context. A stochastic intervention on a set of endogenous variables $X \subseteq B$ assigns a distribution p to X regardless of the other exogenous and endogenous variables as well as the structural functions. We denote by $\mathrm{do}(X \sim p)$ the stochastic intervention on X. An intervention induces a new distribution over the endogenous variables.

For any two variables $X, Y \in B$ with a directed path from X to Y in D, we say the causal effect from X to Y is *confounded* if $p(y|\,\mathrm{do}(X = x)) \neq p(y|X = x)$ (Peters et al., 2017, Def. 6.39).

Appendix C. Proofs

C.1 Proof of Proposition 6

Proof [Proof of Proposition 6]

We recall the trajectories of a behavior policy is generated through Eq. (2.6) with $\{e_h\}_h \perp\!\!\!\perp \{z_h\}_h$. Let $p_{x,h}$ be the marginal distribution of x_h. Also define the probability density function and probability mass function

$$\begin{aligned} p_{a,h}(a \,|\, x, z, e) &:= \pi_{b,h}(a \,|\, x, z, e), \\ p_{x'}(x' \,|\, x, a, e) &:= \mathbb{1}\{x' = F^*(x, a) + e\}. \end{aligned}$$

Then the marginal distribution of $(x_h, a_h, z_h, e_h, x'_h)$, denoted $d_{h,\pi_b,*}$ (we use $*$ to emphasize the presence of unobserved confounder e_h), admits the factorization

$$d_{h,\pi_b,*}(x, a, z, e, x') = \mathcal{P}_z(z)\mathcal{P}_e(e)p_{x,h}(x) \cdot p_{a,h}(a \,|\, x, z, e) \cdot p_{x'}(x' \,|\, x, a, e).$$

And the average visitation distribution of all random variables $\{x_h, a_h, z_h, e_h, x'_h\}_h$ is

$$\begin{aligned} &\bar{d}_{\pi_b,*}(x, a, z, e, x') \\ &:= \frac{1}{H}\sum_{h=1}^{H} d_{h,\pi_b,*}(x, a, z, e, x') \\ &= \mathcal{P}_z(z)\mathcal{P}_e(e) \cdot \left(\sum_{h=1}^{H} p_{x,h}(x)p_{a,h}(a \,|\, x, z, e)\right) \cdot p_{x'}(x' \,|\, x, a, e) \\ &= \mathcal{P}_z(z)\mathcal{P}_e(e) \cdot \left(\frac{1}{H}\sum_{h=1}^{H} p_{x,h}(x)\right) \cdot \left(\sum_{h=1}^{H} \frac{p_{x,h}(x)}{\sum_{k=1}^{H} p_{x,k}(x)} p_{a,h}(a \,|\, x, z, e)\right) \cdot p_{x'}(x' \,|\, x, a, e). \end{aligned}$$

Define the weighted policy $\bar{\pi}(a \,|\, x, z, e) = \big(\sum_{h=1}^{H} p_{x,h}(x)p_{a,h}(a \,|\, x, z, e)\big) / \sum_{h=1}^{H} p_{x,h}(x)$ and the average state visitation distribution $p_x = \frac{1}{H}\sum_{h=1}^{H} p_{x,h}(x)$. Then $(x, a, z, e, x') \sim \bar{d}_{\pi_b,*}$ can be equivalently written as

$$z \sim \mathcal{P}_z,\ e \sim \mathcal{P}_e,\ x \sim p_x,\ a \sim \bar{\pi}(\cdot \,|\, x, z, e),\ x' = F(x, a) + e.$$

We conclude if $(x, a, z, e, x') \sim \bar{d}_{\pi_b,*}$ then $x' = F^*(x, a) + e$ with $\mathbb{E}[e \,|\, z] = 0$.

Remark 16. *We also have $\mathbb{E}[e \,|\, x, z] = 0$ so we could extend the instrument z to $z \leftarrow [x, z]$, and the algorithm and the theory in this paper remain the same.*

■

C.2 Proof of Proposition 7

Proof [Proof of Proposition 7] First note for $f = \phi \cdot \theta \in \mathcal{H}_\phi$, the operator $\Pi_\psi \mathcal{T} f$ admits the form

$$\Pi_\psi \mathcal{T} f = \psi^\top \mathbb{E}[\psi(z)\psi(z)^\top]^{-1}\mathbb{E}[\psi(z)(\theta \cdot \phi)(x,a)] = \psi^\top B^{-1} A\theta.$$

Recall $\|f\|_\phi = \|\theta\|$. Then the feature map ill-poseness can be written as

$$\mu_{\mathrm{IV}} := \min_{f \in \mathcal{H}_\phi} \frac{\|\Pi_\psi \mathcal{T} f\|^2_{L^2(\mathcal{Z})}}{\|f\|^2_\phi} = \min_{\theta \neq 0} \frac{\theta^\top (A^\top B^{-1} A)\theta}{\theta^\top \theta},$$

which is the minimum eigenvalue of the matrix $A^T B^{-1} A$. This completes the proof of Proposition 7. ■

C.3 Proof of Lemma 15

To facilitate the discussion, we recall the definitions of relevant quantities and define some auxiliary operators. We define the operators $\mathbb{J}_h$ and $\widehat{\mathbb{J}}_h$

$$(\mathbb{J}_h f)(x) = \langle f(x,\cdot), \pi^*_h(\cdot \mid x)\rangle, \quad (\widehat{\mathbb{J}}_h f)(x) = \langle f(x,\cdot), \widehat{\pi}_h(\cdot \mid x)\rangle \tag{C.1}$$

for any $h \in [H]$ and function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. For any function $g : \mathcal{S} \to \mathbb{R}$, given the model parameter $\widehat{W}$, define the operator

$$(\widehat{\mathbb{P}} g)(x,a) = \int g(x') \mathcal{P}_{\widehat{W}}(x' \mid x, a)\, \mathrm{d}x',$$

where $\mathcal{P}_{\widehat{W}}(x' \mid x, a)$ is the probability density of d_x-dimensional Gaussian distribution with mean $\widehat{W}\phi(x,a)$ and variance $\sigma^2 I_{d_x}$ (we overload notations and let $\mathcal{P}$ denotes both the distribution and the density of a Gaussian). For the true underlying transition dynamics with model parameter W^*, we define the operator

$$(\mathbb{P} g)(x,a) = \int g(x') \mathcal{P}_{W^*}(x' \mid x, a)\, \mathrm{d}x'. \tag{C.2}$$

We define the quantity

$$\xi_h(x) = (\mathbb{J}_h \widehat{Q}_h)(x) - (\widehat{\mathbb{J}}_h \widehat{Q}_h)(x) = \langle \widehat{Q}_h(x,\cdot), \pi^*_h(\cdot \mid x) - \pi_h(\cdot \mid x)\rangle \tag{C.3}$$

for any $h \in [H]$ and all state $x \in \mathcal{S}$.

Now we clarify the relationship among (π^*, Q^*, V^*), $(\widehat{\pi}, \widehat{Q}, \widehat{V})$ and $(\widehat{\pi}, V^{\widehat{\pi}}, Q^{\widehat{\pi}})$. Recall the Bellman equation of the optimal policy π^*. For $h = 1, \ldots, H$,

$$Q^*_h = r_h + \mathbb{P}(V^*_{h+1}), \quad \forall (x,a), \tag{C.4}$$
$$V^*_h = \langle \pi^*, Q^*_h\rangle = \mathbb{J}_h Q^*_h, \quad \forall x, \tag{C.5}$$
$$V^*_{H+1} = 0 \tag{C.6}$$

and the set of Bellman optimality equations that π^* satisfies: $\pi^*_h(x) = \mathrm{argmax}_a\, Q^*_h(x,a)$, and $V^*_h = \max_a Q^*_h$.

The update rules of $\widehat{\pi}$ in Algorithm 1 imply the following equations relating $\widehat{\pi}$, $\widehat{Q}$ and $\widehat{V}$. For $h = 1, \ldots, H$,

$$\widehat{Q}_h = r_h + \widehat{\mathbb{P}}\widehat{V}_{h+1}, \quad \forall (x,a), \tag{C.7}$$

$$\widehat{\pi}_h(\cdot \,|\, x) = \operatorname*{argmax}_a \widehat{Q}_h(x,a), \quad \forall x, \tag{C.8}$$

$$\widehat{V}_h = \langle \widehat{Q}_h, \widehat{\pi}_h \rangle = \max_a \widehat{Q}_h = \widehat{\mathbb{J}}_h \widehat{Q}_h, \quad \forall x. \tag{C.9}$$

We recall the definition of the model prediction term

$$\iota_h = (r_h + \mathbb{P}\widehat{V}_{h+1}) - \widehat{Q}_h \tag{C.10}$$

for all $(x,a) \in \mathcal{S} \times \mathcal{A}$. Finally, since $Q^{\widehat{\pi}}$ and $V^{\widehat{\pi}}$ are the Q function and value function of the output policy $\widehat{\pi}$, the Bellman equations for $\widehat{\pi}$ holds: for $h = 1, \ldots, H$

$$Q_h^{\widehat{\pi}} = r_h + \mathbb{P}V_{h+1}^{\widehat{\pi}}, \quad \forall (x,a) \tag{C.11}$$

$$V_h^{\widehat{\pi}} = \langle Q_h^{\widehat{\pi}}, \widehat{\pi}_h \rangle = \widehat{\mathbb{J}}_h Q_h^{\widehat{\pi}}, \quad \forall x \tag{C.12}$$

$$V_{H+1}^{\widehat{\pi}} = 0. \tag{C.13}$$

Proof [Proof of Lemma 15] We first write

$$V_1^* - V_1^{\widehat{\pi}} = (V_1^* - \widehat{V}_1) - (\widehat{V}_1 - V_1^{\widehat{\pi}}).$$

Next we analyze the two terms separately.

Part I: Analysis of $(V_1^* - \widehat{V}_1)$. For all state $x \in \mathcal{S}$, and any $h = 1, \ldots, H$

$$V_h^* - \widehat{V}_h = \langle \pi_h^*, Q_h^* \rangle - \langle \widehat{Q}_h, \widehat{\pi}_h \rangle \tag{C.14}$$

$$= \mathbb{J}_h Q_h^* - \widehat{\mathbb{J}}_h \widehat{Q}_h \tag{C.15}$$

$$= \mathbb{J}_h (Q_h^* - \widehat{Q}_h) + (\mathbb{J}_h - \widehat{J}_h)\widehat{Q}_h \tag{C.16}$$

$$= \mathbb{J}_h (Q_h^* - \widehat{Q}_h) + \xi_h \tag{C.17}$$

$$= \mathbb{J}_h ([r_h + \mathbb{P}V_{h+1}^*] - [r_h + \mathbb{P}\widehat{V}_{h+1} - \iota_h]) + \xi_h \tag{C.18}$$

$$= \mathbb{J}_h \mathbb{P}(V_{h+1}^* - \widehat{V}_{h+1}) + \mathbb{J}_h \iota_h + \xi_h. \tag{C.19}$$

Here (C.14) follows from Bellman equations of V_h^* Eq. (C.5) and the update rule of $\widehat{V}_h$ Eq. (C.9); (C.15) follows from the definition of operators $\mathbb{J}_h$ and $\widehat{\mathbb{J}}_h$ Eq. (C.1); in (C.16) we add and subtract $\mathbb{J}_h \widehat{Q}_h$; (C.17) follows from definition of ξ_h in Eq. (C.3); (C.18) follows by using the Bellman equations satisfied by Q_h^* and the definition of ι_h in Eq. (C.10).

Next we apply the above recursion formula for the sequence $\{V_h^* - \widehat{V}_h\}_{h=1}^H$ repeatedly and obtain

$$V_1^* - \widehat{V}_1 = \left(\prod_{h=1}^H \mathbb{J}_h \mathbb{P} \right) \left(V_{H+1}^* - \widehat{V}_{H+1} \right) + \sum_{h=1}^H \left(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P} \right) \mathbb{J}_h \iota_h + \sum_{h=1}^H \left(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P} \right) \xi_h.$$

Using $V_{H+1}^* = \widehat{V}_{H+1} = 0$ gives

$$V_1^* - \widehat{V}_1 = \sum_{h=1}^H \left(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P} \right) \mathbb{J}_h \iota_h + \sum_{h=1}^H \left(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P} \right) \xi_h. \tag{C.20}$$

By definitions of $\mathbb{P}$ in Eq. (C.2), $\mathbb{J}_h$ in Eq. (C.1), and ξ_h in Eq. (C.3), we can equivalently write Eq. (C.20) in the form of expectation w.r.t the optimal policy π^*. For all $x \in \mathcal{S}$,

$$V_1^*(x) - \widehat{V}_1(x) = \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\iota_h(x_h, a_h) \,|\, x_1 = x] + \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\xi_h(x_h)] \,|\, x_1 = x]. \tag{C.21}$$

Part II: Analysis of $(\widehat{V}_1 - V_1^{\widehat{\pi}})$. Notice for any $h = 1, \ldots, H$,

$$\begin{aligned} \widehat{V}_h - V_h^{\widehat{\pi}} &= \widehat{\mathbb{J}}_h \widehat{Q}_h - \widehat{\mathbb{J}}_h Q_h^{\widehat{\pi}} && \text{(C.22)} \\ &= \widehat{\mathbb{J}}_h([r_h + \mathbb{P}\widehat{V}_{h+1} - \iota_h] - [r_h + \mathbb{P}V_h^{\widehat{\pi}}]) && \text{(C.23)} \\ &= \widehat{\mathbb{J}}_h \mathbb{P}(\widehat{V}_{h+1} - V_{h+1}^{\widehat{\pi}}) - \widehat{\mathbb{J}}_h \iota_h. && \text{(C.24)} \end{aligned}$$

Here (C.22) follows from the update rule of $\widehat{V}_h$ Eq. (C.9) and the Bellman equation satisfied by $V_h^{\widehat{\pi}}$ in Eq. (C.12); (C.23) follows from the Bellman equation satisfied by $Q_h^{\widehat{\pi}}$ in Eq. (C.11) and the definition of the model prediction error ι_h in Eq. (C.10).

Apply the recursion repeatedly we obtain

$$\widehat{V}_1 - V_1^{\widehat{\pi}} = \left(\prod_{h=1}^{H} \widehat{\mathbb{J}}_h \mathbb{P}\right) \left(\widehat{V}_{H+1} - V_{H+1}^{\widehat{\pi}}\right) - \sum_{h=1}^{H} \left(\prod_{i=1}^{h-1} \widehat{\mathbb{J}}_i \mathbb{P}\right) \widehat{\mathbb{J}}_h \iota_h$$

Using $\widehat{V}_{H+1} = 0$ by Line 8 of Algorithm 1 and $V_{H+1}^{\widehat{\pi}} = 0$, we obtain

$$\widehat{V}_1 - V_1^{\widehat{\pi}} = -\sum_{h=1}^{H} \left(\prod_{i=1}^{h-1} \widehat{\mathbb{J}}_i \mathbb{P}\right) \widehat{\mathbb{J}}_h \iota_h. \tag{C.25}$$

By definition of $\widehat{\mathbb{J}}_h$ in Eq. (C.1), we write Eq. (C.25) in the form of expectation w.r.t. the policy $\widehat{\pi}$, and we have for all state $x \in \mathcal{S}$

$$\widehat{V}_1(x) - V_1^{\widehat{\pi}}(x) = -\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\iota_h(x_h, a_h) \,|\, x_1 = x]. \tag{C.26}$$

Putting together Eq. (C.21) and Eq. (C.26) completes the proof of Lemma 15. ■

C.4 Proof of Theorem 9

We define

$$\begin{aligned} \mu_A &= \sigma_{\min}\big(\sqrt{A^\top A}\big), & L_A &= \sigma_{\max}\big(\sqrt{A^\top A}\big), \\ \mu_B &= \sigma_{\min}(B), & L_B &= \sigma_{\max}(B)\,, \end{aligned}$$

where for a symmetric positive definite matrix M, the matrix $\sqrt{M}$ is the unique matrix such that $M = \sqrt{M}\sqrt{M}$. Recall the update rule in Algorithm 1 is

$$W_{t+1} = W_t - \eta_t^\theta \cdot (K_t\psi_t)\phi_t^\top, \quad K_{t+1} = K_t + \eta_t^\omega \cdot (K_t\psi_t + x_t' - W_t\phi_t)\psi_t^\top\,. \tag{C.27}$$

Recall the saddle-point problem Eq. (3.6) and we denote the saddle-point function by Φ_i, i.e.,

$$\Phi_i(\theta, \omega) := \theta^\top A^\top \omega - b_i^\top \omega - \tfrac{1}{2}\omega^\top B\omega, \tag{C.28}$$

where $b_i = \mathbb{E}[x_i \psi(z)^\top]$. Given Φ_i defined above, we optimize out the dual variable, and define the primal function P_i and the optimal dual variable $\widehat{\omega}_i$ as follows.

$$P_i(\theta) = \max_{\omega} \Phi_i(\theta, \omega) = \tfrac{1}{2}(A\theta - b_i)^\top B^{-1}(A\theta - b_i) \tag{C.29}$$
$$\widehat{\omega}_i(\theta) = \operatorname*{argmax}_{\omega} \Phi_i(\theta, \omega) = B^{-1}(A\theta - b_i).$$

Uniqueness of $\widehat{\omega}_i(\theta)$ is guaranteed by on the full-rankness of A and B (Assumption A.4). Define by $(\theta_i^{\mathrm{sad}}, \omega_i^{\mathrm{sad}})$ the saddle-point of the convex-concave function Φ_i. Then we have

$$\theta_i^{\mathrm{sad}} = \operatorname*{argmin}_{\theta} P_i(\theta), \quad \omega_i^{\mathrm{sad}} = \widehat{\omega}_i(\theta_i^*).$$

Due to the separable structure of the update Eq. (C.27), if we denote the iterates (W_t, K_t) by $W_t = [\theta_{1,t}, \ldots, \theta_{d_x,t}]^\top$ and $K_t = [\omega_{1,t}, \ldots, \omega_{d_x,t}]^\top$, then we can equivalently write the update as follows. For $i = 1, \ldots, d_x$,

$$\begin{aligned} \theta_{i,t+1} &= \theta_{i,t} - \eta_t^\theta \widetilde{\nabla}_\theta \Phi_i(\theta_{i,t}, \omega_{i,t}) \\ &= \theta_{i,t} - \eta_t^\theta (\phi(x_t, a_t)\psi(z_t)^\top)\omega_{i,t} \qquad \text{(C.30)} \\ \omega_{i,t+1} &= \omega_{i,t} + \eta_t^\omega \widetilde{\nabla}_\omega \Phi_i(\theta_{i,t}, \omega_{i,t}) \\ &= \omega_{i,t} + \eta_t^\omega (\phi(x_t, a_t)^\top \theta_{i,t} - x'_{i,t} - \psi(z_t)^\top \omega_{i,t})\psi(z_t)\,. \qquad \text{(C.31)} \end{aligned}$$

Denote by $(W^{\mathrm{sad}}, K^{\mathrm{sad}})$ the saddle-point of the problem Eq. (3.7). Let $(\theta^{\mathrm{sad}}, \omega^{\mathrm{sad}})$ be the saddle-point of Φ_i in Eq. (C.28). Since the minimax problem Eq. (3.7) is separable in the each coordinate of the primal and the dual variables, we have $\theta^{\mathrm{sad}} = W_i^{\mathrm{sad}}$ and $\omega^{\mathrm{sad}} = K_i^{\mathrm{sad}}$, for all $i = 1, \ldots, d_x$, where W_i^{sad} is the i-th row of the matrix W^{sad}, and K_i^{sad} is the i-the row of K^{sad}. So we turn to study the convergence of $\{\theta_{i,t}, \omega_{i,t}\}_t$ to the saddle-point of Φ_i.

In the rest of the discussion we will ignore the subscript i in $\omega_{i,t}, \theta_{i,t}, x_{i,t}, x'_{i,t}, \Phi_i, P_i, \widehat{\omega}_i$ and b_i. Define the gradient of Φ evaluated at (θ_t, ω_t), $\nabla_\theta \Phi$ and $\nabla_\omega \Phi$, and its stochastic version given a new data tuple $\xi_t = (x_t, a_t, z_t, x'_t)$, $\widetilde{\nabla}_\theta \Phi$ and $\widetilde{\nabla}_\omega \Phi$, by

$$\begin{aligned} &\nabla_\theta \Phi(\theta_t, \omega_t) = A^\top \omega_t, && \widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t; \xi_t) = (\phi(x_t, a_t)\psi(z_t)^\top)\omega_t \qquad \text{(C.32)} \\ &\nabla_\omega \Phi(\theta_t, \omega_t) = A\theta_t - b - B\omega_t, && \widetilde{\nabla}_\omega \Phi(\theta_t, \omega_t; \xi_t) = (\phi(x_t, a_t)^\top \theta_t - x'_t - \psi(z_t)^\top \omega_t)\psi(z_t). \end{aligned}$$

We will ignore the dependence of $\widetilde{\nabla}_\theta \Phi$ and $\widetilde{\nabla}_\omega \Phi$ on ξ_t from now on. Define the auxiliary update sequences given the stochastic update sequence $\{\theta_t, \omega_t\}$ in Eq. (C.30) and Eq. (C.31),

$$\begin{aligned} \widetilde{\theta}_{t+1} &= \theta_t - \eta_t^\theta \nabla_\theta \Phi(\theta_t, \omega_t) && = \theta_t - \eta_t^\theta A^\top \omega_t \\ \widehat{\theta}_{t+1} &= \theta_t - \eta_t^\theta \nabla P(\theta_t) && = \theta_t - \eta_t^\theta A^\top B^{-1}(A\theta_t - b)\,, \\ \widetilde{\omega}_{t+1} &= \omega_t + \eta_t^\omega \nabla \Phi(\theta_t, \omega_t) && = \omega_t + \eta_t^\omega (A\theta_t - b - B\omega_t)\,. \end{aligned}$$

Define the σ-algebras $\mathcal{F}_0 = \sigma\{\theta_0, \omega_0\}$, and $\mathcal{F}_t = \sigma\{\theta_0, \omega_0, \{x_j, a_j, z_j, x'_j\}_{j=0}^{t-1}\}$ for $t = 1, \ldots, T$. Note $\xi_{t-1} \in \mathcal{F}_t$ but $\xi_t \notin \mathcal{F}_t$. Note that for all $t \geq 1$, the random variables ξ_{t-1}, $\theta_t, \omega_t, \widetilde{\theta}_{t+1}, \widetilde{\omega}_{t+1}$ and $\widehat{\theta}_{t+1}$ are deterministic given $\mathcal{F}_t$, and we obviously have

$$\mathbb{E}[\widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t) \,|\, \mathcal{F}_t] = \nabla_\theta \Phi(\theta_t, \omega_t) \quad \text{and} \quad \mathbb{E}[\widetilde{\nabla}_\omega \Phi(\theta_t, \omega_t) \,|\, \mathcal{F}_t] = \nabla_\omega \Phi(\theta_t, \omega_t).$$

We will denote $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \,|\, \mathcal{F}_t]$.

We start with some basic observations of the functions P and Φ.

Lemma 17. *Consider the functions P in Eq.* (C.29) *and Φ in Eq.* (C.28).

1. *Recall μ_{IV} and L_P are the minimum and the maximum eigenvalues of the matrix $A^\top B^{-1} A$, respectively. Then the function P is μ_{IV}-strongly convex and L_P-smooth. Moreover, we have $\mu_{\mathrm{IV}} \geq \mu_A^2 / L_B$, and $L_P \leq \min\{1, L_A^2/\mu_B\}$.*
2. *For any fixed θ, the function $\omega \mapsto -\Phi(\theta, \omega)$ is μ_B-strongly convex and L_B smooth.*
3. *(Proposition 8) Assumptions A.5 and A.4 imply the existence and uniqueness of a matrix $W^* = [W_1^*, \ldots, W_{d_x}^*] \in \mathbb{R}^{d_x \times d_\phi}$ such that $\mathbb{E}[W^* \phi(x,a) \,|\, z] = \mathbb{E}[x' \,|\, z]$. Assumption A.4 implies the uniqueness of the saddle-point $(\theta^{\mathrm{sad}}, \omega^{\mathrm{sad}}) = \operatorname{argmin}_{\theta \in \mathbb{R}^{d_\phi}} \max_{\omega \in \mathbb{R}^{d_\psi}} \Phi_i(\theta, \omega)$. Furthermore, in addition to Assumptions A.5 and A.4, if Assumption A.6 holds, then $W_i^* = \theta^{\mathrm{sad}}$ and $\omega^{\mathrm{sad}} = \widehat{\omega}_i(\widehat{\theta}) = 0$.*

Proof *See Appendix D.1.* ■

Item 3 above shows that under the assumptions listed in Theorem 9, the saddle-point of Φ_i equals to the i-th row of the unknown transition matrix W^*. To emphasize this we now define by (θ^*, ω^*) the saddle-point of the function Φ. Next we present some descent lemmas about the sequence $\{\theta_t, \omega_t\}$. Denote the second moment of the stochastic gradient evaluated at the saddle-point of Φ, (θ^*, ω^*) by

$$\sigma_{\nabla\theta}^2 = \mathbb{E}[\|\widetilde{\nabla}_\theta \Phi(\theta^*, \omega^*)\|^2] \quad \text{and} \quad \sigma_{\nabla\omega}^2 = \mathbb{E}[\|\widetilde{\nabla}_\omega \Phi(\theta^*, \omega^*)\|^2],$$

where $\widetilde{\nabla}_\theta \Phi$ and $\widetilde{\nabla}_\omega \Phi$ are defined in Eq. (C.32). First we show the variance of stochastic gradient can be bounded by the suboptimality of the current iterate.

Lemma 18 (Bounding variance of stochastic gradients)**.** *Consider the sequence $\{\omega_t, \theta_t\}$. If Assumption A.3 holds, then*

$$\begin{aligned}
&\mathbb{E}_t\big[\|\widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t) - \nabla_\theta \Phi(\theta_t, \omega_t)\|^2\big] \\
&\quad \leq 4(\mu_B^{-1} \|\theta_t - \theta^*\|^2 + \|\omega_t - \widehat{\omega}(\theta_t)\|^2) + 2\sigma_{\nabla\theta}^2, \qquad \text{(C.33)} \\
&\mathbb{E}_t\big[\|\widetilde{\nabla}_\omega \Phi(\theta_t, \omega_t) - \nabla_\omega \Phi(\theta_t, \omega_t)\|^2\big] \\
&\quad \leq 16(\mu_B^{-1} \|\theta_t - \theta^*\|^2 + \|\omega_t - \widehat{\omega}(\theta_t)\|^2) + 2\sigma_{\nabla\omega}^2. \qquad \text{(C.34)}
\end{aligned}$$

where we condition on $\mathcal{F}_t$ and take expectation over the new data tuple ξ_t.

Proof See Appendix D.2. ■

Lemma 19 (One-step descent of primal update)**.** *Consider the update sequence $\{\omega_t, \theta_t\}$. Let A.3 (bounded feature map) and A.4 hold. If $\eta_t^\theta \leq \frac{2}{\mu_{\mathrm{IV}} + L_P}$, then*

$$\begin{aligned}
\mathbb{E}\big[\|\theta_{t+1} - \theta^*\|^2\big] \leq\ & (1 - \mu_{\mathrm{IV}} \eta_t^\theta + 4\mu_B^{-1} (\eta_t^\theta)^2) \cdot \mathbb{E}\big[\|\theta_t - \theta^*\|^2\big] \\
& + (\mu_{\mathrm{IV}}^{-1} \eta_t^\theta + 4(\eta_t^\theta)^2) \cdot \mathbb{E}\big[\|\omega_t - \widehat{\omega}(\theta_t)\|^2\big] \\
& + 2(\eta_t^\theta)^2 \cdot \sigma_{\nabla\theta}^2
\end{aligned}$$

Proof See Appendix D.3. ■

Lemma 20 (One-step descent of dual update)**.** *Consider the update sequence* $\{\omega_t, \theta_t\}$*. Let A.3 and A.4 hold. If* $\eta_t^\theta \le \frac{2}{\mu_B + L_B}$*, then*

$$\begin{aligned}\mathbb{E}\big[\|\omega_{t+1} - \widehat{\omega}(\theta_{t+1})\|^2\big] \le& \big(1 - \mu_B\eta_t^\omega + 32(\mu_B^{-2}(\eta_t^\theta)^2(\eta_t^\omega)^{-1} + (\eta_t^\omega)^2 + \mu_B^{-1}(\eta_t^\theta)^2)\big)\cdot \mathbb{E}\big[\|\omega_t - \widehat{\omega}(\theta_t)\|^2\big] \\ &+ 32\big(\mu_B^{-2}(\eta_t^\theta)^2(\eta_t^\omega)^{-1} + \mu_B^{-1}(\eta_t^\omega)^2 + \mu_B^{-2}(\eta_t^\theta)^2\big)\cdot \mathbb{E}\big[\|\theta_t - \theta^*\|^2\big] \\ &+ 32\big((\eta_t^\omega)^2\sigma_{\nabla\omega}^2 + \mu_B^{-1}(\eta_t^\theta)^2\sigma_{\nabla\theta}^2\big)\,. \end{aligned} \tag{C.35}$$

Proof See Appendix D.4. ■

Equipped with Lemmas 19 and 20, we can derive a recursion by choosing appropriate stepsize sequences η_t^ω and η_t^θ. We set

$$\eta_t^\theta = \frac{\beta}{\gamma + t}, \quad \eta_t^\omega = \frac{\alpha\beta}{\gamma + t}$$

for some positive α, β, γ, which will be chosen later. For some positive λ (to be chosen later) we define the potential function P_t with $a_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and $b_t = \mathbb{E}[\|\omega_t - \widehat{\omega}(\theta_t)\|^2]$,

$$P_t = a_t + \lambda b_t,$$

and then derive a recursion formula for P_t. We have by Lemma 19 and 20,

$$\begin{aligned} P_{t+1} &= a_{t+1} + \lambda b_{t+1} \\ &\le \big(1 - \mu_{\mathrm{IV}}\eta_t^\theta + 2^5(\lambda\alpha^{-1}\mu_B^{-2}\eta_t^\theta + \mathrm{III})\big)a_t \\ &+ \big(1 - \mu_B\eta_t^\omega + 2^5(\alpha^{-2}\cdot\mu_B^{-2}\eta_t^\omega + \lambda^{-1}\mu_{\mathrm{IV}}^{-1}\eta_t^\theta + \mathrm{I})\big)(\lambda b_t) \\ &+ \mathrm{II} \end{aligned} \tag{C.36}$$

where

$$\begin{aligned} \mathrm{I} &= (\eta_t^\omega)^2 + \mu_B^{-1}(\eta_t^\theta)^2 + \lambda^{-1}(\eta_t^\theta)^2, \\ \mathrm{II} &= 2(\eta_t^\theta)^2\cdot\sigma_{\nabla\theta}^2 + 4\lambda\big(\mu_B^{-1}(\eta_t^\theta)^2\sigma_{\nabla\theta}^2 + (\eta_t^\omega)^2\sigma_{\nabla\omega}^2\big) \\ \mathrm{III} &= \mu_B^{-1}(\eta_t^\theta)^2 + \lambda\mu_B^{-1}(\eta_t^\omega)^2 + \lambda\mu_B^{-2}(\eta_t^\theta)^2, \end{aligned}$$

Our strategy is straight-forward. We find a suitable choice of the free parameters $(\lambda, \gamma, \alpha, \beta)$ such the the sequence $\widetilde{P}_t$ decays at the rate $1/t$.

Step 1. Choose $\gamma = \gamma(\alpha, \beta, \lambda)$ such that (i) the stepsize requirements in Lemmas 19 and 20 are met, and (ii) the two terms $2^5\cdot\mathrm{III}$ and $2^5\cdot\mathrm{I}$ are less than $\frac{1}{2}\mu_{\mathrm{IV}}\eta_t^\theta$ and $\frac{1}{2}\mu_B\eta_t^\omega$, respectively.

For any positive α, β, λ, we pick γ large enough such that the following inequalities hold for all $t \ge 1$,

$$\begin{aligned} 2^5\cdot\mathrm{III} &\le \frac{1}{2}\mu_{\mathrm{IV}}\eta_t^\theta \\ 2^5\cdot\mathrm{I} &\le \frac{1}{2}\mu_B\eta_t^\omega \end{aligned}$$

Note $\eta_0^\theta = \beta/\gamma$, and $\eta_0^\omega = \alpha\beta/\gamma$. The above inequalities suggest it suffices to set γ large enough. Concretely, for any fixed positive (α, β, λ) with, we can make γ satisfy the following inequalities

$$\gamma \ge 2^8\cdot\max\{\beta\cdot\mu_B^{-1}\mu_{\mathrm{IV}}^{-1}, \alpha^2\lambda\beta\mu_B^{-1}\mu_{\mathrm{IV}}^{-1}, \beta\lambda\mu_B^{-2}\mu_{\mathrm{IV}}^{-1}, \alpha\beta\mu_B^{-1}, \alpha^{-1}\beta\mu_B^{-2}, \alpha^{-1}\lambda^{-1}\beta\mu_B^{-1}\} \tag{C.37}$$

To ensure the stepsizes are small enough to meet the conditions in Lemma 19 and 20 we need for all t,

$$\eta_t^\theta \le \frac{2}{L_P + \mu_{\mathrm{IV}}}, \quad \eta_t^\omega \le \frac{2}{L_B + \mu_B},$$

it suffices to control η_0^θ and η_0^ω by setting

$$\gamma \ge \max\{\beta, \alpha\beta\}. \tag{C.38}$$

For any fixed (α, β, λ), the inequalities Eq. (C.37) and Eq. (C.38) give the choice of γ.

Step 2. Pick α, λ such that the recursion reduces to the form $P_{t+1} \le (1 - \frac{1}{4}\mu_{\mathrm{IV}}\eta_t^\theta)P_t + \text{noise}$. By the choice of γ in Step 1 (Eq. C.37 and Eq. C.38), the recursion Eq. (C.36) reduces to

$$\begin{aligned} P_{t+1} &= a_{t+1} + \lambda b_{t+1} \\ &\le (1 - \frac{1}{2}\mu_{\mathrm{IV}}\eta_t^\theta + 2^5(\lambda\alpha^{-1}\mu_B^{-1}\eta_t^\theta))a_t \\ &+ (1 - \frac{1}{2}\mu_B\eta_t^\omega + 2^5(\alpha^{-2}\cdot\mu_B^{-2}\eta_t^\omega + \lambda^{-1}\mu_{\mathrm{IV}}^{-1}\eta_t^\theta))(\lambda b_t) \\ &+ \mathrm{II} \end{aligned} \tag{C.39}$$

We find (α, λ) such that

$$\begin{aligned} 2^5(\lambda\alpha^{-1}\mu_B^{-2}\eta_t^\theta) &\le \frac{1}{4}\mu_{\mathrm{IV}}\eta_t^\theta \\ 2^5(\alpha^{-2}\cdot\mu_B^{-2}\eta_t^\omega + \lambda^{-1}\mu_{\mathrm{IV}}^{-1}\eta_t^\theta) &\le \frac{1}{4}\mu_B\eta_t^\omega \end{aligned}$$

It suffices to set

$$\lambda = \mu_B^{1/2} \tag{C.40}$$

$$\alpha = 2^8\cdot\mu_B^{-1.5}\mu_{\mathrm{IV}}^{-1} \tag{C.41}$$

Together the choice of λ, α in Eq. (C.40) and Eq. (C.41) implies that the recursion Eq. (C.39) simplifies to

$$P_{t+1} \le \big(1 - \frac{1}{4}\mu_{\mathrm{IV}}\eta_t^\theta\big)a_t + (1 - \frac{1}{4}\mu_B\eta_t^\omega)(\lambda b_t) + \mathrm{II} \tag{C.42}$$

$$\le \big(1 - \frac{1}{4}\mu_{\mathrm{IV}}\eta_t^\theta\big)P_t + \Big(2(\eta_t^\theta)^2\cdot\sigma_{\nabla\theta}^2 + 4\lambda\big(\mu_B^{-1}(\eta_t^\theta)^2\sigma_{\nabla\theta}^2 + (\eta_t^\omega)^2\sigma_{\nabla\omega}^2\big)\Big), \tag{C.43}$$

where we used $1 - \frac{1}{4}\mu_B\eta_t^\omega \le 1 - \frac{1}{4}\mu_{\mathrm{IV}}\eta_t^\theta$ because Eq. (C.41) implies $\alpha \ge \mu_{\mathrm{IV}}\mu_B^{-1}$.

Next we bound the last term in Eq. (C.43). Now we study $\sigma_{\nabla\theta}^2, \sigma_{\nabla\omega}^2$. By Item 3 of Lemma 17, we have the primal variable in the saddle-point of the minimax problem Eq. (C.28) equals to the truth that generates the data, i.e., we have $x_t' = x_{t+1} = (\theta^*)\cdot\phi(x_t, a_t) + e_t$, and that $\omega^* = 0$. The variances of the gradient at the optima (θ^*, ω^*) are

$$\begin{aligned} \sigma_{\nabla\theta}^2 &= \mathbb{E}_{\xi_t}\big[\|\widetilde{\nabla}_\theta\Phi(\theta^*, \omega^*; \xi_t)\|^2\big] \\ &= \mathbb{E}\big[\|(\phi(x_t, a_t)\psi(z_t)^\top)\omega^*\|^2\big] \\ &= 0 \end{aligned}$$

and

$$\begin{aligned}\sigma^2_{\nabla\omega} &= \mathbb{E}_{\xi_t}\big[\|\widetilde{\nabla}_\omega \Phi(\theta^*,\omega^*;\xi_t)\|^2\big] \\ &= \mathbb{E}\big[\|(\phi(x_t,a_t)^\top\theta^* - x_t' - \psi(z_t)^\top\omega^*)\psi(z_t)\|^2\big] \\ &= \mathbb{E}\big[\|e_t\psi(z_t)\|^2\big] \\ &\le \mathbb{E}[e_t^2] = \sigma^2\end{aligned}$$

where we have used $\sup_z \|\psi(z)\|_2 \le 1$ by A.3. This implies

$$2(\eta_t^\theta)^2\cdot\sigma^2_{\nabla\theta} + 4\lambda\big(\mu_B^{-1}(\eta_t^\theta)^2\sigma^2_{\nabla\theta} + (\eta_t^\omega)^2\sigma^2_{\nabla\omega}\big) = \lambda\cdot 4(\eta_t^\omega)^2\cdot\sigma^2.$$

We now restore the omitted state dimension index i, and the recursion Eq. (C.42) writes

$$\begin{aligned}&\mathbb{E}\big[\|\theta_{t+1,i}-\theta_i^*\|^2\big] + \lambda\mathbb{E}\big[\|\omega_{t+1,i} - \widehat{\omega}_i(\theta_{t+1})\|^2\big] \\ &\quad\le \big(1-\frac{1}{4}\mu_{\mathrm{IV}}\eta_t^\theta\big)\Big(\mathbb{E}[\|\theta_{t,i}-\theta_i^*\|^2] + \lambda\mathbb{E}\big[\|\omega_{t,i}-\widehat{\omega}_i(\theta_t)\|^2\big]\Big) + \lambda\cdot 4(\eta_t^\omega)^2\cdot\sigma^2.\end{aligned}$$

Summing over $i = 1,\dots,d_x$, we have a recursion formula on the sequence $\widetilde{P}_t = \mathbb{E}[\|W_t - W^*\|_F^2] + \lambda\mathbb{E}[\|K_t - \widehat{K}(W_t)\|_F^2]$.

$$\widetilde{P}_{t+1} \le (1-\frac{1}{4}\mu_{\mathrm{IV}}\eta_t^\theta)\widetilde{P}_t + \lambda\cdot 4(\eta_t^\omega)^2\cdot d_x\sigma^2. \tag{C.44}$$

Step 3. Pick β,ν such that $\widetilde{P}_t = O(\nu t^{-1})$. Set

$$\begin{aligned}&\beta = 8\mu_{\mathrm{IV}}^{-1}, \\ &\nu = \max\Big\{\gamma\widetilde{P}_0,\ (\frac{1}{4}\mu_{\mathrm{IV}}\beta - 1)^{-1}\beta^2\alpha^2\lambda\cdot d_x\sigma^2\Big\} = \max\{\gamma\widetilde{P}_0,\ \text{const.}\times\mu_{\mathrm{IV}}^{-4}\mu_B^{-2.5}\}.\end{aligned}$$

Together with our choice of α in Eq. (C.41) and λ in Eq. (C.40), we have the following choice of γ (Eq. C.37 and Eq. C.38)

$$\gamma = 2^8\cdot\alpha^2\beta\lambda\cdot\mu_B^{-1}\mu_{\mathrm{IV}}^{-1} = \text{const.}\times\mu_{\mathrm{IV}}^{-4}\mu_B^{-3.5}.$$

Next, we claim for all $t \ge 0$,

$$\widetilde{P}_t \le \frac{\nu}{\gamma+t}. \tag{C.45}$$

We prove by induction. For the base case $t = 0$, the inequality Eq. (C.45) holds by definition of ν. Next, assume for some $t \ge 0$, the inequality Eq. (C.45) holds. We investigate P_{t+1}. By the recursion formula Eq. (C.44),

$$\widetilde{P}_{t+1} \le (1-\frac{1}{4}\mu_{\mathrm{IV}}\eta_t^\theta)\widetilde{P}_t + \lambda\cdot 4(\eta_t^\omega)^2\cdot d_x\sigma^2 \tag{C.46}$$

$$\le \frac{\gamma+t-\frac{1}{4}\mu_{\mathrm{IV}}\beta}{\gamma+t}\cdot\frac{\nu}{\gamma+t} + \lambda\frac{4\alpha^2\beta^2\cdot d_x\sigma^2}{(\gamma+t)^2} \tag{C.47}$$

$$= \frac{(\gamma+t-1)\nu}{(\gamma+t)^2} - \frac{(\frac{1}{4}\mu_{\mathrm{IV}}\beta-1)\nu}{(\gamma+t)^2} + \lambda\frac{4\alpha^2\beta^2\cdot d_x\sigma^2}{(\gamma+t)^2} \tag{C.48}$$

$$\le \frac{\nu}{\gamma+t+1}. \tag{C.49}$$

where Eq. (C.46) holds due to the recursion formula Eq. (C.44); Eq. (C.47) holds due to the induction assumption that $\widetilde{P}_t \le \nu/(\gamma+t)$; Eq. (C.48) holds because (i) $4^{-1}\mu_{\mathrm{IV}}\beta - 1 = 1 \ge 0$ by our choice of β, and (ii) the definition of ν ensures the sum of last two terms in Eq. (C.48) is negative; Eq. (C.49) holds because $(\gamma+t-1)/(\gamma+t)^2 \le (\gamma+t+1)^{-1}$. This proves the claim Eq. (C.45).

This proves Theorem 9.

C.5 Proof of Theorem 9 (ii)

Proof We recall the error decomposition of $V^* - V^{\widehat{\pi}}$ presented in Lemma 15. Conditioning on the training data, the matrix W_T and the functions $\{\iota_h\}_h$ are deterministic. Recall $\xi_h = \langle \widehat{Q}_h, \pi_h^* - \widehat{\pi}_h \rangle_{\mathcal{A}}$ for all $x \in \mathcal{S}$, and $\iota_h = (r_h + \mathbb{P}\widehat{V}_{h+1}) - \widehat{Q}_h$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$. First by definition of $\xi_h = \langle \widehat{Q}_h, \pi_h^* - \widehat{\pi}_h \rangle_{\mathcal{A}}$ and that $\widehat{\pi}_h$ is greedy w.r.t. $\widehat{Q}_h$, we have

$$\sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\xi_h(x_h) \,|\, x_1 = x] \le 0 \quad \text{for all } x.$$

Based on the error decomposition of $V^* - V^{\widehat{\pi}}$ (Lemma 15), we have for all (x,a),

$$\begin{aligned} \|V^* - V^{\widehat{\pi}}\|_\infty &= \sup_x V^*(x) - V^{\widehat{\pi}}(x) \\ &\le \sup_x \left\{ \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\iota_h(x_h,a_h) \,|\, x_1 = x] + \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\iota_h(x_h,a_h) \,|\, x_1 = x] \right\}. \end{aligned} \tag{C.50}$$

Next we derive an upper bound for $\|\iota_h\|_\infty = \sup_{x,a} |\iota_h(x,a)|$.

$$\begin{aligned} \sup_{x,a} |\iota_h(x,a)| &= \sup_{x,a} \left| (r_h + \mathbb{P}\widehat{V}_{h+1}) - \widehat{Q}_h \right| \\ &= \sup_{x,a} \left| (r_h + \mathbb{P}\widehat{V}_{h+1}) - (r_h + \widehat{\mathbb{P}}\widehat{V}_{h+1}) \right| && \text{(C.51)} \\ &= \sup_{x,a} \left| \mathbb{P}\widehat{V}_{h+1} - \widehat{\mathbb{P}}\widehat{V}_{h+1} \right| \\ &\le \sup_{x,a} \left\{ \sqrt{\mathbb{E}_{x' \sim \mathcal{P}_{W^*}(\cdot \,|\, x,a)}\big[\widehat{V}_{h+1}(x')^2\big]} \cdot \min\Big(\frac{\|(W_T - W^*)\phi(x,a)\|_2}{\sigma}, 1 \Big) \right\} && \text{(C.52)} \\ &\le \min\Big\{ \frac{\|W_T - W^*\|}{\sigma}, 1 \Big\} \cdot H. && \text{(C.53)} \end{aligned}$$

Here Eq. (C.51) holds by definition of $\widehat{Q}_h$ Eq. (C.52) holds due to Lemma 24; recall $\mathcal{P}_W(x' \,|\, x, a)$ is the probability density of multivariate Normal with mean $W\phi(x,a)$ and variance $\sigma^2 I_{d_x}$. Eq. (C.53) holds because for all $h \in [H]$ we have $\widehat{V}_h \le H$, and that $\|(W_T - W^*)\phi(x,a)\| \le \|W_T - W^*\| \|\phi(x,a)\|$. Note for all (x,a) we have $\|\phi(x,a)\| \le 1$ (Assumption A.3).

Next we continue from Eq. (C.50).

$$\begin{aligned} \|V^* - V^{\widehat{\pi}}\|_\infty &\le \sup_x \left\{ \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\big[\|\iota_h\|_\infty \,|\, x_1 = x\big] + \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\big[\|\iota_h\|_\infty \,|\, x_1 = x\big] \right\} \\ &\le 2H \cdot \max_{h \in [H]} \|\iota_h\|_\infty \\ &\le 2H^2 \cdot \min\Big\{ \frac{\|W_T - W^*\|}{\sigma}, 1 \Big\} \le 2H^2 \sigma^{-1} \cdot \|W_T - W^*\|. \end{aligned}$$

Now we take expectation on both sides w.r.t. the sampling process, we have

$$\begin{aligned} \mathbb{E}\big[\|V^* - V^{\widehat{\pi}}\|_\infty\big] &\le 2H^2\sigma^{-1} \cdot \mathbb{E}\big[\|W_T - W^*\|\big] \\ &\le 2H^2\sigma^{-1} \cdot \sqrt{\mathbb{E}\big[\|W_T - W^*\|_F^2\big]} \le 2H^2\sigma^{-1} \sqrt{\frac{\nu}{\gamma + T}}\,. \end{aligned}$$

Note we trivially have $\|V^* - V^{\widehat{\pi}}\|_\infty \le H$. So we conclude

$$\mathbb{E}\big[\|V^* - V^{\widehat{\pi}}\|_\infty\big] \le H \cdot \min\left\{2H\sigma^{-1}\sqrt{\frac{\nu}{\gamma+T}}, 1\right\}.$$

This completes the proof of Theorem 9 (ii). ■

C.6 Proof of Theorem 14

Proof [Proof of Theorem 14] Denote $\theta^{\mathrm{sad}} = W_i^{\mathrm{sad}}$. We omit the subscript i in f_i^* and x_i'. This theorem studies the relation between the two quantities:

- An element in the primal function space, $\phi \cdot \theta^* \in \mathcal{H}_\phi$, where θ^* solves the following minimax problem.

$$\min_{f \in \mathcal{H}_\phi} \max_{u \in \mathcal{H}_\psi} \mathbb{E}[(f(x,a) - x')u(z)] - \frac{1}{2}\mathbb{E}[u(z)^2]. \tag{C.54}$$

- The truth f^* that satisfies $\mathbb{E}[f^*(x,a) \,|\, z] = \mathbb{E}[x' \,|\, z]$.

It can be verified that the optimal primal variable of the above minimax problem Eq. (C.54) exists and is unique. Specifically, for $f = \theta \cdot \psi \in \mathcal{H}_\phi$, due to A.4, the inner maximization is uniquely attained at

$$\psi \cdot \widehat{\omega}(\theta) \in \mathcal{H}_\psi, \quad \widehat{\omega}(\theta) := \mathbb{E}[\psi(z)\psi(z)^\top]^{-1}\mathbb{E}\big[\psi(z) \cdot (f(x,a) - x')\big].$$

Also note

$$\psi \cdot \widehat{\omega}(\theta) = \Pi_\psi \mathcal{T}(\theta \cdot \phi - f^*)$$

due to the definition of the projection operator $\Pi_\psi : L^2(\mathcal{Z}) \to \mathcal{H}_\psi$, defined by for all $u \in L^2(\mathcal{Z})$,

$$\Pi_\psi u = \operatorname*{argmin}_{u' \in \mathcal{H}_\psi} \|u - u'\|_{L^2(\mathcal{Z})} = \psi^\top \mathbb{E}[\psi(z)\psi(z)^\top]^{-1}\mathbb{E}[\psi(z)u(z)].$$

Now we plug in the optimal value and define, for $f \in \mathcal{H}_\phi$,

$$\begin{aligned} L(f) &:= \max_{u \in \mathcal{H}_\psi} \mathbb{E}\big[(f(x,a) - x')u(z)\big] - \frac{1}{2}\mathbb{E}[u(z)^2] \\ &= \frac{1}{2}\mathbb{E}[\psi(z) \cdot (f(x,a) - x')]^\top B^{-1}\mathbb{E}[\psi(z) \cdot (f(x,a) - x')] \\ &= \frac{1}{2}\|\Pi_\psi \mathcal{T}(f - f^*)\|^2_{L^2(\mathcal{Z})}. \end{aligned}$$

The unique minimizer of $L(f)$ over $\mathcal{H}_\phi$ is

$$\phi \cdot \theta^{\mathrm{sad}} \in \mathcal{H}_\phi, \quad \theta^{\mathrm{sad}} = [A^\top B^{-1} A]^{-1} A^\top B^{-1}\mathbb{E}[\psi(z)x'] \in \mathbb{R}^{d_\phi}.$$

Note

$$Qf^* = \phi \cdot \theta^{\mathrm{sad}}$$

by definition of the operator Q in Theorem 14. We define $\widehat{f} = \Pi_\phi f^*$, the projection of f^* onto $\mathcal{H}_\phi$ w.r.t the norm $\|\cdot\|_{L^2(\mathcal{S},\mathcal{A})}$. We have the decomposition

$$\|f^* - \theta^{\text{sad}} \cdot \phi\|_{L^2(\mathcal{S},\mathcal{A})} \le \|f^* - \widehat{f}\|_{L^2(\mathcal{S},\mathcal{A})} + \|\widehat{f} - \theta^{\text{sad}} \cdot \phi\|_{L^2(\mathcal{S},\mathcal{A})}.$$

For the first term we have $\|f^* - \widehat{f}\|_{L^2(\mathcal{S},\mathcal{A})} \le \eta_1$ by definition of η_1. For the second term, we further decompose and use the definition of μ_{IV} and Proposition 7.

$$\begin{aligned}
&\|\widehat{f} - \theta^{\text{sad}} \cdot \phi\|_{L^2(\mathcal{S},\mathcal{A})} \\
&\quad\le \|\widehat{f} - \theta^{\text{sad}} \cdot \phi\|_\phi && \text{(C.55)}\\
&\quad\le \mu_{\text{IV}}^{-1} \cdot \|\mathcal{T}(\widehat{f} - \theta^{\text{sad}} \cdot \phi)\|_{L^2(\mathcal{Z})} && \text{(C.56)}\\
&\quad\le \mu_{\text{IV}}^{-1} \cdot \big(\|\mathcal{T}(\widehat{f} - f^*)\|_{L^2(\mathcal{Z})} + \|\mathcal{T}(f^* - \theta^{\text{sad}} \cdot \phi)\|_{L^2(\mathcal{Z})}\big) && \text{(C.57)}\\
&\quad\le \mu_{\text{IV}}^{-1} \cdot \big(\|\mathcal{T}(\widehat{f} - f^*)\|_{L^2(\mathcal{Z})} + \|\Pi_\psi \mathcal{T}(f^* - \theta^{\text{sad}} \cdot \phi)\|_{L^2(\mathcal{Z})} + \eta_2 \cdot \mu\big) && \text{(C.58)}\\
&\quad\le \mu_{\text{IV}}^{-1} \cdot \big(\|\mathcal{T}(\widehat{f} - f^*)\|_{L^2(\mathcal{Z})} + \|\Pi_\psi \mathcal{T}(f^* - \widehat{f})\|_{L^2(\mathcal{Z})} + \eta_2 \cdot \mu\big) && \text{(C.59)}\\
&\quad\le \mu_{\text{IV}}^{-1} \cdot \big(2\|\mathcal{T}(\widehat{f} - f^*)\|_{L^2(\mathcal{Z})} + \eta_2 \cdot \mu\big) && \text{(C.60)}\\
&\quad\le 2c \cdot \eta_1 + \mu_{\text{IV}}^{-1} \cdot \eta_2 \cdot \mu. && \text{(C.61)}
\end{aligned}$$

Here Eq. (C.55) follows since ϕ is bounded; Eq. (C.56) follows by definition of μ_{IV}; Eq. (C.57) follows since T is linear and we use I inequality; Eq. (C.58) follows by definition of η_2 and μ; Eq. (C.59) follows because $\phi^\top \theta^*$ minimizes $f \mapsto \|\Pi_\psi \mathcal{T}(f^* - f)\|^2_{L^2(\mathcal{Z})}$ over $\mathcal{H}_\phi$ and that $\widehat{f} \in \mathcal{H}_\phi$; Eq. (C.60) follows because the projection operator is non-expansive; Eq. (C.61) follows by definition of the constant c.

This completes the proof of Theorem 14. ■

Appendix D. Proofs of Lemmas in Appendix C

D.1 Proof of Lemma 17

Proof

Proof of Item 1 in Lemma 17. For strong convexity, we show that the minimum eigenvalue of $\nabla^2 P(\theta)$ and is lower bounded by $\mu_A^2 L_B^{-1}$. Since the matrix B is full rank (Assumption A.4) and thus its inverse B^{-1} has a unique square root $B^{-1/2}$ such that $B^{-1} = B^{-1/2}B^{-1/2}$. For any $w \in \mathbb{R}^{d_\psi}$ with unit norm we have $\|B^{-1/2}w\| \ge L_B^{-1/2}$. For any $v \in \mathbb{R}^{d_\phi}$ such that $\|v\| = 1$,

$$\begin{aligned}
v^\top \nabla^2 P(\theta) v &= v^\top A^\top B^{-1} A v = v^\top A^\top B^{-1/2} B^{-1/2} A v \\
&= \|B^{-1/2} A v\|^2 \ge L_B^{-1} \|Av\|^2 \ge \mu_A^2 L_B^{-1}
\end{aligned}$$

where we have used the fact that the matrix A has full column rank (Assumption A.4, $\text{rank}(A) = d_\phi$) and thus for any $u \in \mathbb{R}^{d_\phi}$ such that $\|u\| = 1$ we have $\|Au\| \ge \mu_A$. The proof of $L_P \le L_A^2 \mu_B^{-1}$ follows by similar reasoning. To see $L_P \le 1$, recall $D = \mathbb{E}[\phi(x,a)\phi(x,a)^\top]$. We note

$$\|A^\top B^{-1} A\| = \|D^{1/2}(D^{-1/2}A^\top B^{-1/2})(B^{-1/2}A^\top D^{-1/2})D^{1/2}\| \le \|D\| \le 1,$$

where we have used $\|D^{-1/2}A^\top B^{-1/2}\| \le 1$ and by A.3 $\|D\| \le 1$.

Proof of Item 2 in Lemma 17. This is obvious by noting for any θ, $\nabla^2_\omega \Phi(\theta,\omega) = -B$ and that the matrix B satisfies $\mu_B I_{d_\psi} \preceq B \preceq L_B I_{d_\psi}$ with $0 < \mu_B$ (Assumption A.4).

Proof of Item 3 in Lemma 17. The existence of W^* such that $\mathbb{E}[W^*\phi(x,a)\,|\,z] = \mathbb{E}[x'\,|\,x,a]$ is guaranteed by Assumption A.5. From this equation, we multiply both sides by $\mathbb{E}[\phi(x,a)\,|\,z]$ and take expectation w.r.t z, we obtain

$$W\mathbb{E}[\mathbb{E}[\phi(x,a)\,|\,z] \times \mathbb{E}[\phi(x,a)\,|\,z]] = \mathbb{E}[\mathbb{E}[x'\,|\,x,a] \times \mathbb{E}[\phi(x,a)\,|\,z]].$$

So if the matrix $\mathbb{E}[\mathbb{E}[\phi(x,a)\,|\,z] \times \mathbb{E}[\phi(x,a)\,|\,z]]$ is invertible then W^* is the unique solution to the above equation. Such invertibility is implied by Assumption A.4.

Next we show the existence and uniqueness of the saddle-point of Φ_i. For any fixed θ, by full-rankness of B (Assumption A.4), the map $\omega \mapsto \Phi_i(\theta,\omega)$ is uniquely maximized at $\omega = \widehat{\omega}_i(\theta) = B^{-1}(A\theta - b_i)$. Recall $P_i(\theta) = \max_\omega \Phi_i(\theta,\omega) = \frac{1}{2}(A\theta - b_i)^\top B^{-1}(A\theta - b_i)$. By Item 1 of Lemma 17, the minimum eigenvalue of $\nabla^2 P$ is bounded away from zero due to full-rankness of A and B (Assumption A.4). Thus P has a unique minimizer.

Next, we show $W_i^* = \theta^{\text{sad}}$. A.6 implies η_2 in Theorem 14 is zero. A.5 implies η_1 in Theorem 14 is zero. So Theorem 14 implies $W_i^* = \theta^{\text{sad}}$.

Finally we show $\widehat{\omega}_i(\theta^{\text{sad}}) = 0$. Recall $\widehat{\omega}_i(\theta) = B^{-1}(A\theta - b_i)$ for any $\theta \in \mathbb{R}^{d_\phi}$. Recall b_i is defined as $b_i = \mathbb{E}[x_i'\psi(z)]$. Since $\theta^{\text{sad}} = W_i^*$, we have

$$\begin{aligned} A\theta^{\text{sad}} - b_i &= \mathbb{E}[\psi(z)(\phi(x,a)^\top\theta^{\text{sad}} - x_i')] = \mathbb{E}[\psi(z)(\phi(x,a)^\top W_i^* - x_i')] = \mathbb{E}[\psi(z)e_i] \\ &= \mathbb{E}[\psi(z)\mathbb{E}[e_i\,|\,z]] = 0. \end{aligned}$$

We conclude $\widehat{\omega}_i(\theta^{\text{sad}}) = 0$. ■

D.2 Proof of Lemma 18

Proof [Proof of Lemma 18] For the inequality Eq. (C.33), conditioning on $\mathcal{F}_t$, we take expectation over the new data $\xi_t = (x_t, a_t, z_t, x_t' = x_{t+1})$ (note $\xi_t \notin \mathcal{F}_t$)

$$\begin{aligned} \mathbb{E}_t[\|\widetilde{\nabla}_\theta\Phi(\theta_t,\omega_t) - \nabla_\theta\Phi(\theta_t,\omega_t)\|^2] &\le \mathbb{E}_t[\|\widetilde{\nabla}_\theta\Phi(\theta_t,\omega_t)\|^2] \\ &\le 2\mathbb{E}_t[\|\widetilde{\nabla}_\theta\Phi(\theta_t,\omega_t) - \widetilde{\nabla}_\theta\Phi(\theta^*,\omega^*)\|^2] + 2\mathbb{E}_t[\|\widetilde{\nabla}_\theta\Phi(\theta^*,\omega^*)\|^2] \end{aligned}$$

For the first term we use that $\widetilde{\nabla}_\theta\Phi(\theta_t,\omega_t;\xi_t) = (\phi(x_t,a_t)\psi(z_t)^\top)\omega_t$ and that ϕ and ψ are bounded by one (Assumption A.3).

$$\mathbb{E}_t[\|\widetilde{\nabla}_\theta\Phi(\theta_t,\omega_t) - \widetilde{\nabla}_\theta\Phi(\theta^*,\omega^*)\|^2] = \mathbb{E}_t[\|\phi_t\psi_t^\top(\omega_t - \omega^*)\|^2] \le \|\omega_t - \omega^*\|^2$$

We bound $\|\omega_t - \omega^*\|^2$ by

$$\begin{aligned} \|\omega_t - \omega^*\|^2 &\le 2\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 2\|\widehat{\omega}(\theta_t) - \omega^*\|^2 && \text{(D.1)} \\ &= 2\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 2\|(B^{-1}A)(\theta^* - \theta_t)\|^2 \\ &\le 2\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 2L_P\mu_B^{-1} \cdot \|\theta^* - \theta_t\|^2 && \text{(D.2)} \\ &\le 2(\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + \mu_B^{-1} \cdot \|\theta^* - \theta_t\|^2) && \text{(D.3)} \end{aligned}$$

where in Eq. (D.1) we use that $\omega^* = B^{-1}(A\theta^* - b)$ and $\widehat{\omega}(\theta_t) = B^{-1}(A\theta_t - b)$; in Eq. (D.2) we use $\|B^{-1}A\| = \|B^{-1/2}(B^{-1/2}A)\| \le \mu_B^{-1/2}L_P^{-1/2}$; in Eq. (D.3) we use $L_P \le 1$. This completes the proof of the first inequality.

For the second inequality Eq. (C.34) we use similar reasoning.

$$\begin{aligned}
&\mathbb{E}_t[\|\widetilde{\nabla}_\omega \Phi(\theta_t,\omega_t) - \nabla_\omega \Phi(\theta_t,\omega_t)\|^2] \\
&\quad \le \mathbb{E}_t[\|\widetilde{\nabla}_\omega \Phi(\theta_t,\omega_t)\|^2] \\
&\quad \le 2\mathbb{E}_t[\|\widetilde{\nabla}_\omega \Phi(\theta_t,\omega_t) - \widetilde{\nabla}_\omega \Phi(\theta^*,\omega^*)\|^2] + 2\mathbb{E}_t[\|\widetilde{\nabla}_\omega \Phi(\theta^*,\omega^*)\|^2]
\end{aligned}$$

For the first term, note $\widetilde{\nabla}_\omega \Phi(\theta_t,\omega_t;\xi_t) = (\phi(x_t,a_t)^\top \theta_t - x_t' - \psi(z_t)^\top \omega_t)\psi(z_t)$. and thus we have

$$\begin{aligned}
\mathbb{E}_t[\|\widetilde{\nabla}_\omega \Phi(\theta_t,\omega_t) - \widetilde{\nabla}_\omega \Phi(\theta^*,\omega^*)\|^2] &= \mathbb{E}_t[\|\psi_t \phi_t^\top(\theta_t - \theta^*) + \psi_t \psi_t^\top(\omega_t - \omega^*)\|^2] \\
&\le 2\|\theta_t - \theta^*\|^2 + 2\|\omega_t - \omega^*\|^2. \\
&\le (2 + 4L_P \mu_B^{-1})\|\theta_t - \theta^*\|^2 + 4\|\omega_t - \widehat{\omega}(\theta_t)\|^2 \\
&\le 2^3(\mu_B^{-1}\|\theta_t - \theta^*\|^2 + \|\omega_t - \widehat{\omega}(\theta_t)\|^2)
\end{aligned}$$

where we have used A.3, and $\mu_B^{-1} \ge 1$ and $L_P \le 1$. This proves Eq. (C.34). So we complete the proof of Lemma 18. ■

D.3 Proof Lemma 19

Proof [Proof of Lemma 19] Conditioning on $\mathcal{F}_t$, we have

$$\mathbb{E}_t[\|\theta_{t+1} - \theta^*\|^2] = \|\mathbb{E}_t[\theta_{t+1} - \theta^*]\|^2 + \mathbb{E}_t[\|(\theta_{t+1} - \theta^*) - \mathbb{E}_t[\theta_{t+1} - \theta^*]\|^2] \tag{D.4}$$

We bound the first term in Eq. (D.4).

$$\begin{aligned}
\|\mathbb{E}_t[\theta_{t+1} - \theta^*]\|^2 &= \|\widetilde{\theta}_{t+1} - \theta^*\|^2 \\
&\le \big(\|\widehat{\theta}_{t+1} - \theta^*\| + \|\widetilde{\theta}_{t+1} - \widehat{\theta}_{t+1}\|\big)^2 \\
&\le \big((1 - \eta_t^\theta \mu_{\mathrm{IV}})\|\theta_t - \theta^*\| + \|\widetilde{\theta}_{t+1} - \widehat{\theta}_{t+1}\|\big)^2 \qquad \text{(D.5)} \\
&\le (1 - \eta_t^\theta \mu_{\mathrm{IV}})\|\theta_t - \theta^*\|^2 + \frac{1}{\eta_t^\theta \mu_{\mathrm{IV}}}\|\widetilde{\theta}_{t+1} - \widehat{\theta}_{t+1}\|^2, \qquad \text{(D.6)}
\end{aligned}$$

Here in Eq. (D.5) we use Lemma 23 since (i) $\widehat{\theta}_{t+1} = \theta_t - \eta_t^\theta \nabla P(\theta_t)$, (ii) P is μ_{IV}-strongly convex and L_P-smooth (Lemma 17), and (iii) our choice of stepsize. In Eq. (D.6) we use that for any $\epsilon \in (0,1)$, it holds $((1-\epsilon)a + b)^2 \le (1-\epsilon)a^2 + \epsilon^{-1}b^2$; see Lemma 22 for a proof.

We bound the second term in Eq. (D.6) by

$$\begin{aligned}
\big\|\widetilde{\theta}_{t+1} - \widehat{\theta}_{t+1}\big\|^2 &= (\eta_t^\theta)^2 \|\nabla_\theta \Phi(\theta_t,\omega_t) - \nabla_\theta P(\theta_t)\|^2 \\
&= (\eta_t^\theta)^2 \|A^\top \omega_t - A^\top \widehat{\omega}(\theta_t)\|^2 \\
&\le (\eta_t^\theta)^2 L_A^2 \|\omega_t - \widehat{\omega}(\theta_t)\|^2 .
\end{aligned}$$

Continuing from Eq. (D.6), we have

$$\big\|\mathbb{E}_t[\theta_{t+1} - \theta^*]\big\|^2 \le (1 - \eta_t^\theta \mu_{\mathrm{IV}})\|\theta_t - \theta^*\|^2 + \eta_t^\theta \cdot L_A^2 \mu_{\mathrm{IV}}^{-1} \cdot \|\omega_t - \widehat{\omega}(\theta_t)\|^2 \tag{D.7}$$

Next we bound the second term in Eq. (D.4).

$$\begin{aligned}
\mathbb{E}_t\big[\big\|(\theta_{t+1} - \theta^*) - \mathbb{E}_t[\theta_{t+1} - \theta^*]\big\|^2\big] &= \mathbb{E}_t\big[\big\|\theta_{t+1} - \mathbb{E}_t[\theta_{t+1}]\big\|^2\big] \\
&= \mathbb{E}_t\big[\big\|\theta_{t+1} - \widetilde{\theta}_{t+1}\big\|^2\big] \\
&= (\eta_t^\theta)^2 \cdot \mathbb{E}_t\big[\big\|\widetilde{\nabla}_\theta \Phi(\theta_t,\omega_t) - \nabla \Phi(\theta_t,\omega_t)\big\|^2\big] . \qquad \text{(D.8)}
\end{aligned}$$

This can be bounded by Lemma 18. Plugging into Eq. (D.4) the bounds in Eq. (D.7) and Eq. (D.8),

$$\begin{aligned}\mathbb{E}_t[\|\theta_{t+1}-\theta^*\|^2] &\le (1-\eta_t^\theta\mu_{\mathrm{IV}})\|\theta_t-\theta^*\|^2+(\eta_t^\theta)L_A^2\mu_{\mathrm{IV}}^{-1}\|\omega_t-\widehat{\omega}(\theta_t)\|^2\\ &\quad+(\eta_t^\theta)^2\mathbb{E}_t\big[\big\|\widetilde{\nabla}_\theta\Phi(\theta_t,\omega_t)-\nabla\Phi(\theta_t,\omega_t)\big\|^2\big]\\ &\le (1-\eta_t^\theta\mu_{\mathrm{IV}})\|\theta_t-\theta^*\|^2+(\eta_t^\theta)L_A^2\mu_{\mathrm{IV}}^{-1}\big\|\omega_t-\widehat{\omega}(\theta_t)\big\|^2\\ &\quad+(\eta_t^\theta)^2\cdot\big(4\big\|\omega_t-\widehat{\omega}(\theta_t)\big\|^2+4L_P\mu_B^{-1}\|\theta_t-\theta^*\|^2+2\sigma_{\nabla\theta}^2\big)\end{aligned}$$

where we have used Lemma 18. Taking expectation on both sides, we get

$$\begin{aligned}\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2] &\le \big(1-\mu_{\mathrm{IV}}\eta_t^\theta+4L_P\mu_B^{-1}(\eta_t^\theta)^2\big)\cdot\mathbb{E}\big[\|\theta_t-\theta^*\|^2\big]\\ &\quad+\big(L_A^2\mu_{\mathrm{IV}}^{-1}\eta_t^\theta+4(\eta_t^\theta)^2\big)\cdot\mathbb{E}\big[\big\|\omega_t-\widehat{\omega}(\theta_t)\big\|^2\big]\\ &\quad+2(\eta_t^\theta)^2\cdot\sigma_{\nabla\theta}^2\\ &\le \big(1-\mu_{\mathrm{IV}}\eta_t^\theta+4\mu_B^{-1}(\eta_t^\theta)^2\big)\cdot\mathbb{E}\big[\|\theta_t-\theta^*\|^2\big]\\ &\quad+\big(\mu_{\mathrm{IV}}^{-1}\eta_t^\theta+4(\eta_t^\theta)^2\big)\cdot\mathbb{E}\big[\big\|\omega_t-\widehat{\omega}(\theta_t)\big\|^2\big]\\ &\quad+2(\eta_t^\theta)^2\cdot\sigma_{\nabla\theta}^2\end{aligned}$$

where we use $L_P\le 1$ and $L_A\le 1$. This completes the proof of Lemma 19. ■

D.4 Proof of Lemma 20

Proof [Proof of Lemma 20] We first bound the one-step difference of primal updates.

Lemma 21 (One-step difference)**.** *Consider the update sequence $\{\omega_t,\theta_t\}$. Conditioning on $\mathcal{F}_t$, we have*

$$\big\|\mathbb{E}_t[\theta_{t+1}-\theta_t]\big\|^2\le 2(\eta_t^\theta)^2\big(L_P^2\cdot\|\theta_t-\theta^*\|^2+L_A^2\cdot\|\omega_t-\widehat{\omega}(\theta_t)\|^2\big)\,.$$

Proof [Proof of Lemma 21] We start by noting

$$\begin{aligned}\big\|\mathbb{E}_t[\theta_{t+1}-\theta_t]\big\|^2&=\big\|\widetilde{\theta}_{t+1}-\theta_t\big\|^2=(\eta_t^\theta)^2\cdot\|A^\top\omega_t\|^2\\ &\le(\eta_t^\theta)^2\cdot\big(2\|A^\top\widehat{\omega}(\theta_t)\|^2+2\|A^\top\omega_t-A^\top\widehat{\omega}(\theta_t)\|^2\big)\,.\end{aligned}\tag{D.9}$$

For the first term in Eq. (D.9), we have

$$\|A^\top\widehat{\omega}(\theta_t)\|=\|\nabla P(\theta_t)\|=\|\nabla P(\theta_t)-\nabla P(\theta^*)\|\le L_P\|\theta_t-\theta^*\|.\tag{D.10}$$

For the second term in Eq. (D.9), we have

$$\|A^\top\omega_t-A^\top\widehat{\omega}(\theta_t)\|^2\le L_A^2\|\omega_t-\widehat{\omega}(\theta_t)\|^2.\tag{D.11}$$

Plugging into Eq. (D.9) the bounds in Eq. (D.10) and Eq. (D.11), we complete the proof of Lemma 21. ■

Now we prove Lemma 20. Conditioning on $\mathcal{F}_t$, we have

$$\mathbb{E}_t\big[\|\omega_{t+1}-\widehat{\omega}(\theta_{t+1})\|^2\big]=\big\|\mathbb{E}_t[\omega_{t+1}-\widehat{\omega}(\theta_{t+1})]\big\|^2\tag{D.12}$$

$$+\,\mathbb{E}_t\Big[\big\|\big(\omega_{t+1}-\widehat{\omega}(\theta_{t+1})\big)-\mathbb{E}_t\big[\omega_{t+1}-\widehat{\omega}(\theta_{t+1})\big]\big\|^2\Big]\,.\tag{D.13}$$

Next we bound the first term in Eq. (D.12)

$$\begin{aligned}
\big\|\mathbb{E}_t\big[\omega_{t+1}-\widehat{\omega}(\theta_{t+1})\big]\big\|^2 &= \big\|\mathbb{E}_t\big[\omega_{t+1}-\widehat{\omega}(\theta_t)\big]+\mathbb{E}_t\big[\widehat{\omega}(\theta_t)-\widehat{\omega}(\theta_{t+1})\big]\big\|^2 \\
&\le \Big(\big\|\mathbb{E}_t\big[\omega_{t+1}-\widehat{\omega}(\theta_t)\big]\big\|+\big\|\mathbb{E}_t\big[\widehat{\omega}(\theta_t)-\widehat{\omega}(\theta_{t+1})\big]\big\|\Big)^2 \\
&= \Big(\|\widetilde{\omega}_{t+1}-\widehat{\omega}(\theta_t)\|+\big\|\mathbb{E}_t\big[\widehat{\omega}(\theta_t)-\widehat{\omega}(\theta_{t+1})\big]\big\|\Big)^2 \\
&\le \Big((1-\mu_B\eta_t^\omega)\|\omega_t-\widehat{\omega}(\theta_t)\|+\big\|\mathbb{E}_t\big[B^{-1}A(\theta_t-\theta_{t+1})\big]\big\|\Big)^2 \qquad \text{(D.14)} \\
&\le \Big((1-\mu_B\eta_t^\omega)\|\omega_t-\widehat{\omega}(\theta_t)\|+L_A\mu_B^{-1}\cdot\big\|\mathbb{E}_t[\theta_t-\theta_{t+1}]\big\|\Big)^2 \\
&\le (1-\mu_B\eta_t^\omega)\|\omega_t-\widehat{\omega}(\theta_t)\|^2+L_P\mu_B^{-1}\cdot\frac{1}{\mu_B\eta_t^\omega}\cdot\big\|\mathbb{E}_t[\theta_t-\theta_{t+1}]\big\|^2. \qquad \text{(D.15)}
\end{aligned}$$

Here in Eq. (D.14) we use that (i) $\widetilde{\omega}_{t+1}=\omega_t+\eta_t^\omega\nabla\Phi(\theta_t,\omega_t)$, (ii) for θ_t, the map $\omega\mapsto-\Phi(\theta_t,\omega)$ is μ_B-strongly convex and L_B-smooth (Lemma 17), (iii) our choice of stepsize, and (iv) $\widehat{\omega}(\theta_t)$ is the minimizer of the map $\omega\mapsto-\Phi(\theta_t,\omega)$. In Eq. (D.15) we use that for any $\epsilon\in(0,1)$, any $a,b\in\mathbb{R}$, it holds $((1-\epsilon)a+b)^2\le(1-\epsilon)a^2+\epsilon^{-1}b^2$.

Using Lemma 21 we can bound the second term in Eq. (D.15) by $\|\omega_t-\widehat{\omega}(\theta_t)\|$ and $\|\theta_t-\theta^*\|$.

Now we bound the second term in Eq. (D.13).

$$\begin{aligned}
&\mathbb{E}_t\big[\|(\omega_{t+1}-\widehat{\omega}(\theta_{t+1}))-\mathbb{E}_t[\omega_{t+1}-\widehat{\omega}(\theta_{t+1})]\|^2\big] \\
&\quad\le \mathbb{E}_t\Big[2\big\|\omega_{t+1}-\mathbb{E}_t[\omega_{t+1}]\big\|^2+2\big\|\widehat{\omega}(\theta_{t+1})-\mathbb{E}_t\big[\widehat{\omega}(\theta_{t+1})\big]\big\|^2\Big]. \qquad \text{(D.16)}
\end{aligned}$$

For the first term in Eq. (D.16) we have

$$\begin{aligned}
\mathbb{E}_t\big[\big\|\omega_{t+1}-\mathbb{E}_t[\omega_{t+1}]\big\|^2\big] &= (\eta_t^\omega)^2\cdot\mathbb{E}_t\big[\big\|\widetilde{\nabla}_\omega\Phi(\theta_t,\omega_t)-\nabla_\omega\Phi(\theta_t,\omega_t)\big\|^2\big] \\
&\le (\eta_t^\omega)^2\cdot\big(16\|\theta_t-\theta^*\|^2+16\|\omega_t-\widehat{\omega}(\theta_t)\|^2+2\sigma_{\nabla\omega}^2\big)
\end{aligned}$$

where we have used Lemma 18. For the second term in Eq. (D.16) we have

$$\begin{aligned}
&\mathbb{E}_t\big[\big\|\widehat{\omega}(\theta_{t+1})-\mathbb{E}_t[\widehat{\omega}(\theta_{t+1})]\big\|^2\big] \\
&\quad=\mathbb{E}_t\big[\big\|B^{-1}A\theta_{t+1}-\mathbb{E}_t[B^{-1}A\theta_{t+1}]\big\|^2\big] \\
&\quad\le L_P\mu_B^{-1}\cdot\mathbb{E}_t\big[\|\theta_{t+1}-\mathbb{E}_t[\theta_{t+1}]\|^2\big] \\
&\quad= L_P\mu_B^{-1}\cdot(\eta_t^\theta)^2\cdot\mathbb{E}_t\big[\big\|\widetilde{\nabla}_\theta\Phi(\theta_t,\omega_t)-\nabla_\theta\Phi(\theta_t,\omega_t)\big\|^2\big] \\
&\quad\le \mu_B^{-1}\cdot(\eta_t^\theta)^2\cdot(4\|\omega_t-\widehat{\omega}(\theta_t)\|^2+4\mu_B^{-1}\|\theta_t-\theta^*\|^2+2\sigma_{\nabla\theta}^2). \qquad \text{(D.17)}
\end{aligned}$$

where we have used Lemma 18 in Eq. (D.17).

Continuing from Eq. (D.16) (the variance part), we obtain

$$\begin{aligned}
&\mathbb{E}_t\Big[\big\|\big(\omega_{t+1}-\widehat{\omega}(\theta_{t+1})\big)-\mathbb{E}_t[\omega_{t+1}-\widehat{\omega}(\theta_{t+1})]\big\|^2\Big] \\
&\quad\le 2^5\big(\mu_B^{-1}(\eta_t^\omega)^2+\mu_B^{-2}(\eta_t^\theta)^2\big)\|\theta_t-\theta^*\|^2 \\
&\qquad+2^5\big((\eta_t^\omega)^2+\mu_B^{-1}(\eta_t^\theta)^2\big)\|\omega_t-\widehat{\omega}(\theta_t)\|^2 \\
&\qquad+4\big(\mu_B^{-1}(\eta_t^\theta)^2\sigma_{\nabla\theta}^2+(\eta_t^\omega)^2\sigma_{\nabla\omega}^2\big) \qquad \text{(D.18)}
\end{aligned}$$

Putting together Eq. (D.15), Eq. (D.18) and Lemma 21, we have

$$\begin{aligned}\mathbb{E}_t[\|\omega_{t+1}-\widehat{\omega}(\theta_{t+1})\|^2] \leq & (1-\mu_B\eta_t^\omega)\|\omega_t-\widehat{\omega}(\theta_t)\|^2 \\ & + 2^5\big(\mu_B^{-2}(\eta_t^\theta)^2/\eta_t^\omega + \mu_B^{-1}(\eta_t^\omega)^2 + \mu_B^{-2}(\eta_t^\theta)^2\big)\|\theta_t-\theta^*\|^2 \\ & + 2^5\big(\mu_B^{-2}(\eta_t^\theta)^2/\eta_t^\omega + (\eta_t^\omega)^2 + \mu_B^{-1}(\eta_t^\theta)^2\big)\|\omega_t-\widehat{\omega}(\theta_t)\|^2 \\ & + 4\big(\mu_B^{-1}(\eta_t^\theta)^2\sigma_{\nabla\theta}^2 + (\eta_t^\omega)^2\sigma_{\nabla\omega}^2\big)\,. \end{aligned} \tag{D.19}$$

This completes the proof of Lemma 20 ■

Appendix E. Supporting Lemmas

Lemma 22. *For any $\epsilon \in (0,1)$, any $a, b \in \mathbb{R}$, it holds $((1-\epsilon)a+b)^2 \leq (1-\epsilon)a^2 + \epsilon^{-1}b^2$.*

Proof By the Cauchy–Schwarz inequality, we have for all $\beta > 0$, $(a+b)^2 \leq (1+\beta)a^2 + (1+\beta^{-1})b^2$. Setting $\beta = \epsilon(1-\epsilon)^{-1}$ completes the proof. ■

Lemma 23 (One-step gradient descent for smooth and strongly-convex function)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a β-smooth and α-strongly convex function. Let $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. For any $0 < \eta \leq \frac{2}{\alpha+\beta}$ and any $x \in \mathbb{R}^d$, let $x^+ = x - \eta\nabla f(x)$. Then $\|x^+ - x^*\| \leq (1-\alpha\eta)\|x-x^*\|$.*

Proof See Lemma 3.1 of Du and Hu (2019). ■

Lemma 24 (Expectation Difference Under Two Gaussians, Lemma C.2 in Kakade et al. 2020)**.** *For Gaussian distribution $\mathcal{N}(\mu_1, \sigma^2 I)$ and $\mathcal{N}(\mu_2, \sigma^2 I)$ ($\sigma^2 \neq 0$), for any positive measurable function g, we have*

$$\mathbb{E}_{z\sim N_1}[g(z)] - \mathbb{E}_{z\sim N_2}[g(z)] \leq \min\left\{\frac{\|\mu_1-\mu_2\|}{\sigma}, 1\right\}\sqrt{\mathbb{E}_{z\sim\mathcal{N}_1}[g(z)^2]}\,.$$

Proof For completeness we present a proof. Note

$$\begin{aligned}\mathbb{E}_{z\sim N_1}[g(z)] - \mathbb{E}_{z\sim N_2}[g(z)] &= \mathbb{E}_{z\sim N_1}\left[g(z)\left(1-\frac{\mathcal{N}_2(z)}{\mathcal{N}_1(z)}\right)\right] \\ &\leq \sqrt{\mathbb{E}_{z\sim N_1}[g(z)^2]}\sqrt{\int\frac{(\mathcal{N}_1(z)-\mathcal{N}_2(z))^2}{\mathcal{N}_1(z)}dz} \\ &= \sqrt{\mathbb{E}_{z\sim N_1}[g(z)^2]}\sqrt{\exp\left(\frac{|\,\mu_1-\mu_2\|^2}{2\sigma^2}\right)-1}\,.\end{aligned}$$

Since $g \geq 0$ we have $\mathbb{E}_{z\sim N_1}[g(z)] - \mathbb{E}_{z\sim N_2}[g(z)] \leq \mathbb{E}_{z\sim N_1}[g(z)] \leq \sqrt{\mathbb{E}_{z\sim N_1}[g(z)^2]}$. Finally, we use $\exp(x) \leq 1+2x$ for $0 \leq x \leq 1$.

$$\begin{aligned}\mathbb{E}_{z\sim N_1}[g(z)] - \mathbb{E}_{z\sim N_2}[g(z)] &\leq \sqrt{\mathbb{E}_{z\sim N_1}[g(z)^2]}\sqrt{\min\left\{\exp\left(\frac{\|\mu_1-\mu_2\|^2}{2\sigma^2}\right)-1, 1\right\}} \\ &\leq \sqrt{\mathbb{E}_{z\sim N_1}[g(z)^2]}\cdot\min\left\{\frac{\|\mu_1-\mu_2\|}{\sigma}, 1\right\}\,.\end{aligned}$$

This completes the proof. ■

References

Donald W.K. Andrews. Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics*, 199:213–220, 2017.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in Artificial Intelligence*, 1994.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed Markov decision processes. *Operations Research*, 2023.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, 2019.

Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Blair Bilodeau, Linbo Wang, and Dan Roy. Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems*, 2022.

Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75(6):1613–1669, 2007.

M. Alan Brookhart and Sebastian Schneeweiss. Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *The International Journal of Biostatistics*, 3(1), 2007.

M Alan Brookhart, Til Stürmer, Robert J Glynn, Jeremy Rassen, and Sebastian Schneeweiss. Confounding control in healthcare database research: Challenges and potential approaches. *Medical Care*, 48:114–120, 2010.

Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 2020.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6:5633–5751, 2007.

Bibhas Chakraborty and Erica E.M. Moodie. *Statistical Methods for Dynamic Treatment Regimes*. Springer, 2013.

Bibhas Chakraborty and Susan A. Murphy. Dynamic treatment regimes. *Annual Review of Statistics and Its Application*, 1(1):447–464, 2014.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

Shuxiao Chen and Bo Zhang. Estimating and improving dynamic treatment regimes with a time-varying instrumental variable. *Journal of the Royal Statistical Society: Series B*, 85(2):427–453, 2023.

Xiaohong Chen and Timothy M Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84, 2018.

Victor Chernozhukov and Christian Hansen. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Victor Chernozhukov, Guido W Imbens, and Whitney K Newey. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4–14, 2007.

Yifan Cui and Eric Tchetgen Tchetgen. A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association*, 116(533): 162–173, 2021.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, 2018.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

Xavier D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27:460–471, 2011.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, 2020.

Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, 2017.

Zuyue Fu, Zhengling Qi, Zhaoran Wang, Zhuoran Yang, Yanxun Xu, and Michael R Kosorok. Offline reinforcement learning with instrumental variables in confounded Markov decision processes. *arXiv preprint arXiv:2209.08666*, 2022.

Joseph Futoma, Anthony Lin, Mark Sendak, Armando Bedoya, Meredith Clement, Cara O'Brien, and Katherine Heller. Learning to treat sepsis with multi-output Gaussian process deep recurrent Q-networks. 2018.

Arthur Guez, Robert D Vincent, Massimo Avoli, and Joelle Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

Joel L Horowitz and Sokbae Lee. Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75(4):1191–1208, 2007.

Yingyao Hu and Ji-Liang Shiu. Nonparametric identification using instrumental variables: Sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693, 2017.

Paul Hünermund and Elias Bareinboim. Causal inference and data fusion in econometrics. *The Econometrics Journal*, 2023.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, 2020.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, 2018.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

Rainer Kress. *Linear Integral Equations*. Springer, 1989.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

Luofeng Liao, You-Lin Chen, Yang Zhuoran, Bo Dai, Mladen Kolar, and Zhaoran Wang. Provably efficient neural estimation of structural equation model: An adversarial approach. In *Advances in Neural Information Processing Systems*, 2020.

S. A. Lorch, M. Baiocchi, C. E. Ahlberg, and D. S. Small. The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics*, 130(2):270–278, 2012.

Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *Journal of Machine Learning Research*, 23(32):1–30, 2022.

Charles F Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Haben Michael, Yifan Cui, Scott A Lorch, and Eric J Tchetgen Tchetgen. Instrumental variable estimation of marginal structural mean models for time-varying treatment. *Journal of the American Statistical Association*, pages 1–12, 2023.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual IV: A single stage instrumental variable regression. In *Advances in Neural Information Processing Systems*, 2020.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B*, 65(2):331–355, 2003.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, 2019.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Advances in Neural Information Processing Systems*, 2020.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Lam Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter R chtarik, Katya Scheinberg, and Martin Takac. SGD and hogwild! Convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, 2018.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *International Conference on Machine Learning*, 2019.

Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Combining kernel and model based learning for HIV therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239–248, 2017.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

Hongming Pu and Bo Zhang. Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society: Series B*, 83(2): 318–345, 2021.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, 2017.

Tongzheng Ren, Tianjun Zhang, Csaba Szepesvári, and Bo Dai. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*, 2022.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

P Rosenbaum. *Observational Studies*. Springer, 2002.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *AAAI Conference on Artificial Intelligence*, 2020.

Jialei Wang and Lin Xiao. Exploiting strong convexity from data with primal-dual first-order algorithms. In *International Conference on Machine Learning*, 2017.

Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B*, 80(3):531–550, 2018.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. In *Advances in Neural Information Processing Systems*, 2021.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommender systems. In *ACM Conference on Recommender Systems*, 2020.

Yang Xu, Jin Zhu, Chengchun Shi, Shikai Luo, and Rui Song. An instrumental variable approach to confounded off-policy evaluation. In *International Conference on Machine Learning*, 2023.

Mengxin Yu, Zhuoran Yang, and Jianqing Fan. Strategic decision-making in the presence of information asymmetry: Provably efficient RL with algorithmic instruments. *arXiv preprint arXiv:2208.11040*, 2022.

Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical Report R-23, Columbia CausalAI Laboratory, 2016.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, 2019.

Junzhe Zhang and Elias Bareinboim. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, 2020.

Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcomes. In *AAAI Conference on Artificial Intelligence*, 2021.