
Instrumental Variable Value Iteration for Causal Offline Reinforcement Learning

Luofeng Liao *
University of Chicago

Zuyue Fu *
Northwestern University

Zhuoran Yang
Princeton University

Yixin Wang
University of California, Berkeley

Mladen Kolar
University of Chicago

Zhaoran Wang
Northwestern University

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 are all mediated through 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 (CMR) 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 CMR. To the best of our knowledge, this is the first provably efficient algorithm for instrument-aided offline RL.

1 Introduction

In reinforcement learning (RL) [64], 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 [60, 37, 27, 50, 28, 55, 58].

One of the major concerns of working with observational data, especially for RL applications in healthcare, is confounding caused by unobserved variables. Because the available data may not contain measurements of important prognostic variables that guide treatment decisions, or heuristic information such as visual inspection of or discussions with patients during each treatment period, variables that affect both the treatment decisions and the next-stage health status of patients are present. See [11] for a detailed discussion of sources of confounding in healthcare datasets.

Instrumental variables (IVs) are a very 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 only affects Y through X . IVs are commonly used in healthcare studies 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 the preference-based IVs (see Example A.1), distance to a specialty care provider, and genetic variants [3]. We introduce one example below.

Example 1.1 (Differential travel time as IV, NICU application). [41, 46, 18] study the effect on neonatal mortality of delivery at high-level neonatal intensive care units (NICU), using the same differential travel time as IV. The goal is to design a neonatal regionalization system that designates

*Equal contribution

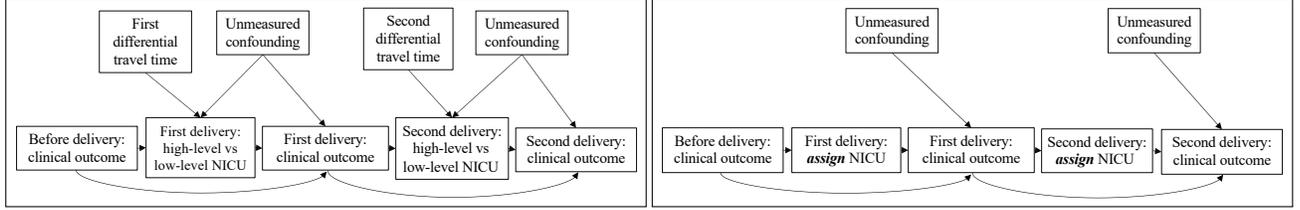


Figure 1: The NICU application, adapted from [18, Figure 1]. Left panel: DAG representing data generation process where UCs are present. Right panel: DAG representing a prenatal regionalization system in action.

hospitals according to the level of care infants need. The available dataset has $\sim 180,000$ 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 mothers’ choice of NICU and does not impact clinical outcomes through other means. A neonatal regionalization system (Figure 1, right panel) designates NICU solely based on the clinical outcome at the previous stage (since differential travel time does not affect clinical outcome anymore once we actually *assign* NICU, and confounders remain unobserved), removing arrows pointing into NICU decision in the DAG presented in the upper panel.

We summarize three aspects of offline medical datasets often encountered by RL practitioners: (i) there is a large amount of logged data where the actual effects of medical treatment on patient’s health are confounded, (ii) the potential presence of a valid IV has been argued for in the biostatistics and epidemiology literature, and (iii) it is expensive or unethical to do experimentation and then inspect the actual performance of a target treatment 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 we termed 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 $O(\mu_{IV}^{-4} \mu_B^{-2.5} H^4 d_x \epsilon^{-2})$, where $0 < \mu_{IV} < 1$ 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 sample complexity result for an IV-aided offline RL.

1.1 Related Work

RL in the presence of UCs has attracted increasing attention; see §A for a detailed overview of related work. 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 [62], 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 sequence of work in [36, 34, 35, 50]. 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 (2.6)), instrumental variables can

enable **point identification** of the structural quantity through conditional moment restriction (along with certain completeness assumptions; see Remark C.1), 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 [59, 18]. These works, and in particular [18], rely on partial identification bounds in the fully nonparametric IV setting [44, 4]. 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.

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 $[n] = \{1, \dots, n\}$. For any bounded function $\varphi: \mathcal{X} \rightarrow \mathbb{R}^{d_\varphi}$, we define the linear function space spanned by φ as $\mathcal{H}_\varphi = \{\theta \cdot \varphi: \theta \in \mathbb{R}^{d_\varphi}\}$. 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 §2.1 as a preliminary. In §2.2.1, we describe the *evaluation setting*, where we test the performance of our learned policy. In §2.2.2, we describe the *offline 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 offline setting.

2.1 Preliminaries: Instrumental Variables

We define confounders and IVs as follows.

Definition 2.1 (Confounders and Instrumental Variables, [56]). *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 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 offline 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 §2.1 to the multi-stage decision making process. In §B we discuss possible extension of this model.

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 $\mathcal{U} \subseteq \mathbb{R}^{d_u}$ is the space of UCs; the integer H is the length of each episode; and $r = \{r_h: \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]\}_{h=1}^H$ 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 $F^*: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ 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 [64]: 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

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

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

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

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

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

Here, the expectation \mathbb{E}_π 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 (2.2) for any $i \in \{h, h+1, \dots, H\}$.

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

$$\|V_1^* - V_1^\pi\|_\infty := \sup_{x \in \mathcal{S}} V_1^*(x) - V_1^\pi(x). \quad (2.4)$$

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 $\langle Q_h^\pi(x, \cdot), \pi_h(\cdot | x) \rangle_{\mathcal{A}} = \int_{\mathcal{A}} Q_h^\pi(x, a) \pi_h(da | x)$ 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 $f : \mathcal{S} \rightarrow \mathbb{R}$. The subscript \mathcal{A} is omitted subsequently if it is clear from the context. Similarly, the Bellman optimality equation takes the following form,

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

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.

2.2.2 Offline Setting: Data Collection Process

We describe the offline 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 $\{e_h\}_h \stackrel{\text{iid}}{\sim} \mathcal{P}_e$, and a sequence of observable IVs $\{z_h\}_h \stackrel{\text{iid}}{\sim} \mathcal{P}_z$. 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,

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

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 $h \in [H]$.

A causal DAG is given in Figure 2 (left) 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

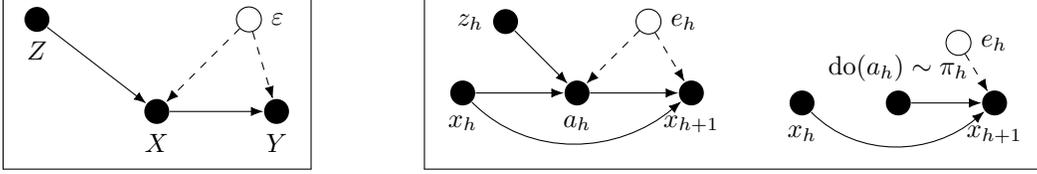


Figure 2: Left panel: An illustration of Definition 2.1 with one UC ε and three observable variables X , Y , and Z . Right panel: *Offline setting* of CMDP-IV with a behavior policy π_b (left). *Evaluation setting* of CMDP-IV with intervention induced by π (right).

(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 (2.6).

The main difference between the evaluation setting (2.1) and the offline setting (2.6) is whether the UC e_h has an effect on the action a_h . In the language of causal inference [56], a policy $\pi = \{\pi_h \in \Delta(\mathcal{A}; \mathcal{S})\}_{h=1}^H$ induces the stochastic intervention $\text{do}(a_1 \sim \pi_1(\cdot | x_1), \dots, a_H \sim \pi_H(\cdot | x_H))$ 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). §E includes more details on the do-operation.

Under the offline 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 (2.6) in the offline setting, can we find a policy that minimizes the suboptimality defined in (2.4) in the evaluation setting?

The challenge of the problem stems from the fact that the UC e_h enters both of the equations (2.6). In general we do not have $\mathbb{E}[x_{h+1} | x_h, a_h] = F^*(x_h, a_h)$ in the offline dynamics; see Remark B.3.

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 §3.1.2.

3.1.1 Conditional Moment Restriction

Following the confounded dynamics (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,

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

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

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

The following proposition states the conditional moment restriction (CMR) implied by the IVs in the offline confounder dynamics (2.6). See §F.1 for the proof.

Proposition 3.1 (CMR). *If (x, a, z, x') is distributed according to the law \bar{d}_{π_b} , then for any $z \in \mathcal{Z}$,*

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

Proposition 3.1 implies that the transition function F^* satisfies the equation $\mathcal{T}F^* = \mathbb{E}[x' | z]$, where the operator \mathcal{T} is defined in (3.2). Such an equation is a Fredholm integral equation of the first kind [38]. Given data collected from \bar{d}_{π_b} , we aim to estimate F^* based on the CMR.

Remark 3.2 (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.

Our model can also be extended to the case of global UCs, assuming that the global UCs have the same effect on the states x_h in both offline and evaluation settings. Only their effects on the actions a_h can differ; the global UCs affect the actions offline but not in evaluation settings. In more detail, the global UCs 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. Yet, assuming that the effect of global UCs on the states are persistent in both evaluation and offline settings, our results would extend to global UCs. Moreover, with additional assumptions on the transition dynamics, some settings of global UCs can be reduced to our setting. 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 §2.1 to the multi-stage decision making process.

3.1.2 A Primal-Dual Estimand

We derive a primal-dual estimand for $F^* = [f_1^*, \dots, f_{d_x}^*]^\top$. For any $i \in [d_x]$, by Proposition 3.1, $\mathbb{E}[f_i^*(x, a) | z] = \mathbb{E}[x'_i | z]$, where x'_i is the i -th element of the next state x' . We find f_i^* by solving the least-square problem $\min_{f_i \in L^2(\mathcal{S}, \mathcal{A})} \frac{1}{2} \mathbb{E}[(\mathbb{E}[f_i(x, a) | z] - \mathbb{E}[x'_i | z])^2]$. By Fenchel duality, the least-square problem admits a primal-dual formulation

$$\min_{f_i \in L^2(\mathcal{S}, \mathcal{A})} \max_{u_i \in L^2(\mathcal{Z})} \mathbb{E}[(f_i(x, a) - x'_i)u_i(z)] - \frac{1}{2} \mathbb{E}[u_i(z)^2], \quad (3.4)$$

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

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

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

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

where the expectations are taken following \bar{d}_{π_b} . We replace the L^2 spaces in (3.4) by their finite-dimensional subspaces, $\min_{f_i \in \mathcal{H}_\phi} \max_{u_i \in \mathcal{H}_\psi} \mathbb{E}[(f_i(x, a) - x'_i)u_i(z)] - \frac{1}{2} \mathbb{E}[u_i(z)^2]$, which, in matrix form, writes

$$\min_{\theta_i \in \mathbb{R}^{d_\phi}} \max_{\omega_i \in \mathbb{R}^{d_\psi}} \omega_i^\top A \theta_i - b_i \omega_i - \frac{1}{2} \omega_i^\top B \omega_i, \quad (3.6)$$

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

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

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

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.1.
 - 2: **Phase 1 (Estimation of W^{sad} in Eq. (3.7))**
 - 3: **for** $t = 0, 1, \dots, T - 1$ **do**
 - 4: $\phi_t \leftarrow \phi(x_t, a_t), \psi_t \leftarrow \psi(z_t)$.
 - 5: $W_{t+1} \leftarrow W_t - \eta_t^\theta \cdot (K_t \psi_t \phi_t^\top), \quad K_{t+1} \leftarrow K_t + \eta_t^\omega \cdot (K_t \psi_t \psi_t^\top + x'_t x'_t{}^\top - W_t \phi_t \psi_t^\top)$.
 - 6: **end for**
 - 7: **Phase 2 (Value iteration)**
 - 8: $\widehat{V}_{H+1}(\cdot) \leftarrow 0, \quad \widehat{W} \leftarrow W_T$.
 - 9: **for** $h = H, H - 1, \dots, 1$ **do**
 - 10: $\widehat{Q}_h(\cdot, \cdot) \leftarrow r_h(\cdot, \cdot) + \int_{\mathcal{S}} \widehat{V}_{h+1}(x') \mathcal{P}_{\widehat{W}}(dx' | \cdot, \cdot)$.
 - 11: $\widehat{\pi}_h(\cdot) \leftarrow \operatorname{argmax}_a \widehat{Q}_h(\cdot, a), \quad \widehat{V}_h(\cdot) \leftarrow \max_a \widehat{Q}_h(\cdot, a)$.
 - 12: **end for**
 - 13: **Output:** $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$.
-

3.2 Algorithm

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

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

Assumption A.1 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 offline setting. Algorithm 1 consists of the following two phases.

Phase 1. In Lines 3–7 of Algorithm 1, we solve (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 (3.5), gives us the updates of W_{t+1} and K_{t+1} in Line 5, respectively.

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 (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 $f: \mathcal{S} \rightarrow \mathbb{R}$.

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

Assumption A.3 (Nondegenerate feature maps). *It holds that $\operatorname{rank}(A) = d_\phi$ and $\operatorname{rank}(B) = d_\psi$ for A and B defined in (3.5).*

Uniqueness of W^{sad} . Assumption A.3 implies the minimax problem (3.7) admits a unique solution. In the min-max problem (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 $\max_K L(W, K) = \frac{1}{2} \operatorname{Tr}[(WA^\top - C)B^{-1}(WA^\top - C)^\top]$. 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.3 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 [2]. For RL applications in healthcare, 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,

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

where Π_ψ is the projection operator onto the space \mathcal{H}_ψ , i.e., $\Pi_\psi u = \operatorname{argmin}_{u' \in \mathcal{H}_\psi} \|u - u'\|_{L^2(\mathcal{Z})}^2$ for any $u \in L^2(\mathcal{Z})$. The definition of μ_{IV} in (4.1) mimics the notion of *sieve measure of ill-posedness* well-known in the literature on NPIV as a measure of IV strength [9, 19]. We next show μ_{IV} admits a simple expression.

Proposition 4.1. *Let A.3 hold. Then $\mu_{IV} = \sigma_{\min}(A^\top B^{-1}A)$.*

4.1 Parametric Case

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

Assumption A.4 (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 [32, 43] in the context of dynamical system identification.

Assumption A.5 (Realizability). *For all $f \in \mathcal{H}_\phi$, it holds that $\mathcal{T}f \in \mathcal{H}_\psi$.*

Proposition 4.2. *Let A.3, A.4 and A.5 hold. Then $W^* = W^{\text{sad}}$.*

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 4.3 in §D. The complete proofs are given in §F.4.

Theorem 4.3 (Parametric case). *Let A.4–A.5 hold. We set the stepsizes in Algorithm 1 as $\eta_t^\theta = \beta/(\gamma + t)$ and $\eta_t^\omega = \alpha \eta_t^\theta$ for any $t \in [T]$, where $\alpha = c_1 \mu_{IV}^{-1} \mu_B^{-1.5}$, $\beta = c_2 \mu_{IV}^{-1}$, $\gamma = c_3 \mu_{IV}^{-4} \mu_B^{-3.5}$, and c_1, c_2, c_3 are positive absolute constants. Then*

(i) *the estimation error satisfies*

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

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

(ii) *the planning error satisfies*

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

The expectation is taken over the data.

For an appropriately chosen initial estimates W_0 and K_0 , Theorem 4.3 shows that the sample complexity needed to recover an ϵ -optimal policy using offline data is of order

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

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 4.4 (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 C.2. From an optimization perspective, the saddle-point problem (3.6) is a stochastic convex-strongly-concave one, a case rarely investigated in the optimization literature; see Remark C.3 for a brief review. 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.

Remark 4.5 (Dependence on IV strength). In (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 $O(\mu_{IV}^{-4} \mu_B^{-2.5} \sigma^2 T^{-1})$. 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 [24] under A.5. There the proposed estimator is the saddle-point of the sample version of (3.6); see Remark C.2 for more details. In particular, they provide a bound in the L^2 -norm, and the order of the variance term is $O(\mu_{IV}^{-4} \max\{d_\phi, d_\psi\} T^{-1})$ ². The minimax optimal rate for NPIV problem is established in the work of [9], attained by sieve estimators. In comparison, the variance term in the minimax optimal rate is of order $O(\widetilde{\mu}_{IV}^{-2} d_\psi T^{-1})$ ³, where $\widetilde{\mu}_{IV}$ is the minimum nonzero singular value of $D^{-1/2} A B^{-1/2}$, quantifying the strength of an IV in a similar way to our μ_{IV} .

Remark 4.6 (Dependence on horizon and state dimension). The work of [32] 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 [32]. Despite that we deal with confounders in additive nonlinear dynamics, our dependence on d_x and H matches their sample complexity bounds.

4.2 Nonparametric Case

In A.4 and A.5 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 our theory to the nonparametric case (e.g., $F^* : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^{d_x}$ is Hölder continuous, and functions of the form $\{\mathcal{T}f \mid f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}, \text{ bounded and continuous}\}$ are also Hölder continuous), we need to discuss two issues. The first one is identification: whether F^* is the unique solution to the CMR (3.3). Identification in NPIV usually requires some form of completeness assumptions; see Remark C.1. The second issue is the error caused by finite-dimensional approximation which we address below.

Let f^* be one element of $F^* = [f_1^*, \dots, f_{d_x}^*]^\top$. If A.4 is violated, we define the primal approximation error $\eta_1 := \|f^* - \Pi_\phi f^*\|_{L^2(\mathcal{S}, \mathcal{A})}$. If A.5 is violated, we define the dual error, which characterizes how well the dual function space \mathcal{H}_ψ approximates functions of the form $\mathcal{T}(f - f^*)$ for $f \in \mathcal{H}_\phi$. Formally we define $\eta_2 := \sup\{\|\mathcal{T}f - \Pi_\psi \mathcal{T}f\|_{L^2(\mathcal{Z})} : f \in \mathcal{H}_\phi, \|f\|_{L^2(\mathcal{S}, \mathcal{A})} \leq 1\}$. Obviously A.4 implies $\eta_1 = 0$ and A.5 implies $\eta_2 = 0$.

We show that, when A.4 and A.5 are violated, the difference between F^* and $W^{\text{sad}}\phi$ 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 $F^* = [f_1^*, \dots, f_{d_x}^*]^\top$, and W_i^{sad} is the i -th row of the estimand W^{sad} defined in (3.7).

Theorem 4.7 (Nonparametric case). *Let A.3 hold. Assume there is a constant $c > 0$ such that $\mu_{IV}^{-1} \cdot \|\mathcal{T}(f_i^* - \Pi_\phi f_i^*)\|_{L^2(\mathcal{Z})} \leq c \cdot \|f_i^* - \Pi_\phi f_i^*\|_{L^2(\mathcal{S}, \mathcal{A})}$. We define the operator $Q : L^2(\mathcal{S}, \mathcal{A}) \mapsto \mathcal{H}_\phi$, $Qf = \operatorname{argmin}_{f' \in \mathcal{H}_\phi} \|\Pi_\psi \mathcal{T}(f' - f)\|_{L^2(\mathcal{Z})}$. Let $\mu = \|(\Pi_\phi - Q)f_i^*\|_{L^2(\mathcal{S}, \mathcal{A})}$. It holds*

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

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 Conclusion

Our model is motivated by real-world applications of RL in healthcare, where it is often the case that 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 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.

²In Appendix D of [24], their (γ_n, k_n, m_n) is the same as our $(\mu_{IV}^{-1}, d_\phi, d_\psi)$.

³In Theorem 2 of [9], their (τ_n, k_n) is the same as our $(\widetilde{\mu}_{IV}^{-1}, d_\phi)$.

References

- [1] D. W. Andrews. Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics*, 199:213–220, 2017.
- [2] J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist’s companion*. Princeton University Press, 2008.
- [3] M. Baiocchi, J. Cheng, and D. S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.
- [4] A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in Artificial Intelligence*, 1994.
- [5] A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.
- [6] E. Bareinboim and J. Pearl. Causal inference by surrogate experiments: z -identifiability. In *Conference on Uncertainty in Artificial Intelligence*, 2012.
- [7] A. Bennett, N. Kallus, and T. Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*. 2019.
- [8] A. Bennett, N. Kallus, L. Li, and A. Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. *arXiv preprint arXiv:2007.13893*, 2020.
- [9] R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75(6):1613–1669, 2007.
- [10] M. A. Brookhart and S. Schneeweiss. Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *The International Journal of Biostatistics*, 3(1), Jan. 2007.
- [11] M. A. Brookhart, T. Stürmer, R. J. Glynn, J. Rassen, and S. Schneeweiss. Confounding control in healthcare database research: Challenges and potential approaches. *Medical Care*, 48: 114–120, 2010.
- [12] L. Buesing, T. Weber, Y. Zwols, S. Racaniere, A. Guez, J.-B. Lespiau, and N. Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.
- [13] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 2020.
- [14] M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6: 5633–5751, 2007.
- [15] B. Chakraborty and E. E. Moodie. *Statistical Methods for Dynamic Treatment Regimes*. Springer, 2013. doi: 10.1007/978-1-4614-7428-9. URL <https://doi.org/10.1007/978-1-4614-7428-9>.
- [16] B. Chakraborty and S. A. Murphy. Dynamic treatment regimes. *Annual Review of Statistics and Its Application*, 1(1):447–464, Jan. 2014. doi: 10.1146/annurev-statistics-022513-115553. URL <https://doi.org/10.1146/annurev-statistics-022513-115553>.
- [17] A. Chambolle and T. 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.
- [18] S. Chen and B. Zhang. Estimating and improving dynamic treatment regimes with a time-varying instrumental variable. *arXiv preprint arXiv:2104.07822*, 2021.
- [19] X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84, 2018.

- [20] B. Dai, N. He, Y. Pan, B. Boots, and L. Song. Learning from conditional distributions via dual embeddings. In *International Conference on Artificial Intelligence and Statistics*, 2017.
- [21] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, 2018.
- [22] S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.
- [23] X. D’Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27:460–471, 2011.
- [24] N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis. Minimax estimation of conditional moment models. *arXiv preprint arXiv:2006.07201*, 2020.
- [25] S. S. Du and W. 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.
- [26] S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, 2017.
- [27] J. Futoma, A. Lin, M. Sendak, A. Bedoya, M. Clement, C. O’Brien, and K. Heller. Learning to treat sepsis with multi-output Gaussian process deep recurrent Q-networks. 2018. URL <https://openreview.net/forum?id=SyxCqGbrZ>.
- [28] A. Guez, R. D. Vincent, M. Avoli, and J. Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.
- [29] Y. Hu and J.-L. Shiu. Nonparametric identification using instrumental variables: Sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693, June 2017.
- [30] P. Hünermund and E. Bareinboim. Causal inference and data-fusion in econometrics. *arXiv preprint arXiv:1912.09104*, 2019.
- [31] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.
- [32] S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi, and W. Sun. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, 2020.
- [33] N. Kallus and A. Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, 2018.
- [34] N. Kallus and A. Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *arXiv preprint arXiv:2002.04518*, 2020.
- [35] N. Kallus and A. Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, May 2021. doi: 10.1287/mnsc.2020.3699. URL <https://doi.org/10.1287/mnsc.2020.3699>.
- [36] N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.
- [37] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.
- [38] R. Kress. *Linear Integral Equations*. Springer, 1989.

- [39] G. Lewis and V. Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.
- [40] L. Liao, Y.-L. Chen, Y. Zhuoran, B. Dai, M. Kolar, and Z. Wang. Provably efficient neural estimation of structural equation model: An adversarial approach. In *Advances in Neural Information Processing Systems*. 2020.
- [41] 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, July 2012. doi: 10.1542/peds.2011-2820. URL <https://doi.org/10.1542/peds.2011-2820>.
- [42] C. Lu, B. Schölkopf, and J. M. Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.
- [43] H. Mania, M. I. Jordan, and B. Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.
- [44] C. F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.
- [45] W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.
- [46] H. Michael, Y. Cui, S. Lorch, and E. T. Tchetgen. Instrumental variable estimation of marginal structural mean models for time-varying treatment. *arXiv preprint arXiv:2004.11769*, 2020.
- [47] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual IV: A single stage instrumental variable regression. In *Advances in Neural Information Processing Systems*, 2020.
- [48] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, Apr. 2003. doi: 10.1111/1467-9868.00389. URL <https://doi.org/10.1111/1467-9868.00389>.
- [49] O. Nachum, Y. Chow, B. Dai, and L. Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, 2019.
- [50] H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 2020.
- [51] 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.
- [52] W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.
- [53] L. Nguyen, P. H. Nguyen, M. van Dijk, P. R. Chahar, K. Scheinberg, and M. Takac. SGD and hogwild! Convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, 2018.
- [54] M. Oberst and D. Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *International Conference on Machine Learning*, 2019.
- [55] S. Parbhoo, J. Bogojeska, M. Zazzi, V. Roth, and F. Doshi-Velez. Combining kernel and model based learning for HIV therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239–248, 2017.
- [56] J. Pearl. *Causality*. Cambridge University Press, 2009.
- [57] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

- [58] N. Prasad, L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.
- [59] H. Pu and B. Zhang. Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):318–345, Mar 2021. ISSN 1467-9868. doi: 10.1111/rssb.12413. URL <http://dx.doi.org/10.1111/rssb.12413>.
- [60] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi. Continuous state-space models for optimal sepsis treatment: A deep reinforcement learning approach. *arXiv preprint arXiv:1705.08422*, 2017.
- [61] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 09 1951.
- [62] P. Rosenbaum. *Observational Studies*. Springer, 2002.
- [63] R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, 2019.
- [64] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.
- [65] G. Tennenholtz, S. Mannor, and U. Shalit. Off-policy evaluation in partially observable environments. *arXiv preprint arXiv:1909.03739*, 2019.
- [66] J. Wang and L. Xiao. Exploiting strong convexity from data with primal-dual first-order algorithms. In *International Conference on Machine Learning*, 2017.
- [67] J. Zhang and E. Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical Report R-23, Columbia CausalAI Laboratory, 2016.
- [68] J. Zhang and E. Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, 2019.
- [69] J. Zhang and E. Bareinboim. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, 2020.
- [70] J. Zhang and E. Bareinboim. Bounding causal effects on continuous outcomes. In *AAAI Conference on Artificial Intelligence*, 2021.

A More Related Work

Dynamic treatment regime (DTR) DTR [48, 15, 16] is a popular model for sequential decision making. DTR 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. A concurrent work by [18] study the policy improvement problem in the presence of UCs, using partial identification results of causal quantities with IVs [44, 4]. 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. [67] formulate the MDP with unobserved confounding using the language of structural causal models. [42] study a model-based RL algorithm in a combined online and offline setting. They propose a structural causal model for the confounded MDP and estimate the structural function with neural nets using the observational data. [12] propose a model-based RL algorithm in the evaluation setting that learns the optimal policy for a partially observable Markov decision process (POMDP). [54] 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. [65] study offline policy evaluation in POMDP. Their identification strategy relies on the identification results of proxy variables in causal inference [45]. [68, 69] 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. [50] 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. [8] 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. [34] study off-policy evaluation in infinite horizon. They characterize the partially identified set of policy values and compute bounds on such a set. [33, 35] 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 [52, 14, 19, 22], and very recently, reproducing kernel Hilbert space-based estimators [63, 47]. 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 [20, 39, 7, 47, 24, 40].

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 [61]. 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 C.2 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.

Example A.1 (Preference-based IV, MIMIC-III data). For example, the work of [10] 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 [37] on the MIMIC-III dataset [31], 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 .

B Appendix to §2

Remark B.1 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 [8] where the authors study policy evaluation with unobserved confounders, confounders are also assumed i.i.d.

Remark B.2 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 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 [6, 30]. 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 [4, 5, 70].

Remark B.3 (The challenge of UCs). The challenge stems from the fact that the UC e_h enters both of the equations (2.6). For ease of discussion, suppose that the behavior policy π_b is deterministic. With slight abuse of notations, we denote by $\pi_{b,h} : \mathcal{S} \times \mathcal{Z} \times \mathcal{U} \rightarrow \mathcal{A}$ the deterministic behavior policy at the h -th step for any $h \in [H]$. Now, (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 $\pi_{b,h}^{-1} : \mathcal{S} \times \mathcal{Z} \times \mathcal{A} \rightarrow \mathcal{U}$. Then, by substituting $e_h = \pi_{b,h}^{-1}(x_h, z_h, a_h)$ into (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.

C Appendix to §4

Remark C.1 On completeness conditions. Bounded completeness condition is a relatively weak regularity assumption on the average visitation distribution \bar{d}_{π_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, [9, 23, 29, 1].

In the parametric case ($F^* = W^* \phi$, for some known bounded feature map ϕ), bounded completeness is more than enough to ensure identification. In fact, it suffices to impose invertibility on the matrix $\mathbb{E}[\mathbb{E}[\phi(x, z) | z] \times \mathbb{E}[\phi(x, z) | z]]$ to ensure uniqueness and existence of the matrix W^* .

Remark C.2 Stochastic approximation for instrumental variables. Our stochastic approximation (SA) estimation procedure is in contrast with the empirical saddle-point estimator proposed in [24]. To estimate f_j^* , their estimator would be defined as the solution to the finite-sum saddle-point problem

$$\operatorname{argmin}_{f \in \mathcal{H}_\phi} \max_{u \in \mathcal{H}_\psi} \frac{1}{n} \sum_{i=1}^n \left\{ (f(x_i, a_i) - x'_{i,j})u(z_i) + \frac{1}{2}u(z_i)^2 \right\} - \frac{\lambda}{2} \|u\|_\phi^2 + \frac{\mu}{2} \|f\|_{\mathcal{H}_\psi}^2 \quad (\text{C.1})$$

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}_{π_b} , and $x'_{i,j}$ denotes the j -th coordinate of $x'_i \in \mathbb{R}^{d_x}$. Their procedure faces two challenges: (i) using the correct regularization parameter, and (ii) finding an approximate solution of the convex-concave optimization problem

(C.1), 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 in related primal-dual methods in RL; see, e.g., [20, 21, 49]. In contrast, the SA approach considered in this work tackles computational error and statistical error jointly and enjoys a fast rate of $O(1/T)$.

Remark C.3 More on $O(1/T)$ rate. We now review literature that studies convex-strongly-concave SSP. A slow rate $O(1/\sqrt{T})$ is obvious by the results for general stochastic convex-concave problem [51]. The work of [17] studies deterministic CSC problem with bilinear coupling and shows a $O(1/T^2)$ rate. [66, 26, 25] 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.

D 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 4.4 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

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

where $K^*(W) = (WA^\top - C)B^{-1}$ with A , B and C defined in (3.5). The matrix $K^*(W)$ is the optimal dual variable in the saddle-point problem (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 [51], we follow the idea in the work of [53] where we upper bounds the variance of stochastic gradients by the suboptimality of the current iterate; see Lemma F.3. Thus our algorithm does not require projection in each iteration. A careful analysis of the recursion for the sequence $\{\tilde{P}_t\}$ 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 [13].

Lemma D.1 (Suboptimality Decomposition). *It holds that for all states $x \in \mathcal{S}$,*

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

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

Proof. See Appendix F.3 for a detailed proof. \square

The output policy $\{\hat{\pi}_h\}$ is greedy with respect to the Q-functions $\{\hat{Q}_h\}$, and therefore the second term on the right-hand side of (D.1) is negative. The model prediction error term ι_h quantifies the mismatch of the pair $(\hat{V}_{h+1}, \hat{Q}_h)$ as the solution to the Bellman equation. This term is controlled by bounding the chi-squared distance between two normal distributions with means $W^*\phi(x, a)$ and $\hat{W}_T\phi(x, a)$, respectively.

E 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 [56, 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 $X_j = f_j(X_{\text{pa}_D(j)}, U_j)$ for some $f_j \in F, U_j \in B$, where $\text{pa}_D(j)$ 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 $\text{do}(X = x)$ the intervention on X and write $\text{do}(x)$ 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 $\text{do}(X \sim p)$ 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 $p(y | \text{do}(X = x)) \neq p(y | X = x)$ [57, Def. 6.39].

F Proofs

F.1 Proof of Proposition 3.1

Proof of Proposition 3.1. We recall the trajectories of a behavior policy is generated through (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 $\bar{\pi}(a | x, z, e) = (\sum_{h=1}^H p_{x,h}(x) p_{a,h}(a | x, z, e)) / \sum_{h=1}^H p_{x,h}(x)$ 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 F.1. 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. □

E.2 Proof of Proposition 4.1

Proof of Proposition 4.1. 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-posedness can be written as

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

which is the minimum eigenvalue of the matrix $A^\top B^{-1}A$. This completes the proof of Proposition 4.1. \square

E.3 Proof of Lemma D.1

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

$$(\mathbb{J}_h f)(x) = \langle f(x, \cdot), \pi_h^*(\cdot | x) \rangle, \quad (\widehat{\mathbb{J}}_h f)(x) = \langle f(x, \cdot), \widehat{\pi}_h(\cdot | x) \rangle \quad (\text{F.1})$$

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

$$(\widehat{\mathbb{P}}g)(x, a) = \int g(x') \mathcal{P}_{\widehat{W}}(x' | x, a) dx',$$

where $\mathcal{P}_{\widehat{W}}(x' | x, a)$ 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

$$(\mathbb{P}g)(x, a) = \int g(x') \mathcal{P}_{W^*}(x' | x, a) dx'. \quad (\text{F.2})$$

We define the quantity

$$\xi_h(x) = (\mathbb{J}_h \widehat{Q}_h)(x) - (\widehat{\mathbb{J}}_h \widehat{Q}_h)(x) = \langle \widehat{Q}_h(x, \cdot), \pi_h^*(\cdot | x) - \widehat{\pi}_h(\cdot | x) \rangle \quad (\text{F.3})$$

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 $h = 1, \dots, H$,

$$Q_h^* = r_h + \mathbb{P}(V_{h+1}^*), \quad \forall (x, a), \quad (\text{F.4})$$

$$V_h^* = \langle \pi^*, Q_h^* \rangle = \mathbb{J}_h Q_h^*, \quad \forall x, \quad (\text{F.5})$$

$$V_{H+1}^* = 0 \quad (\text{F.6})$$

and the set of Bellman optimality equations that π^* satisfies: $\pi_h^*(x) = \operatorname{argmax}_a Q_h^*(x, a)$, 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 $h = 1, \dots, H$,

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

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

$$\widehat{V}_h = \langle \widehat{Q}_h, \widehat{\pi}_h \rangle = \max_a \widehat{Q}_h = \widehat{\mathbb{J}}_h \widehat{Q}_h, \quad \forall x. \quad (\text{F.9})$$

We recall the definition of the model prediction term

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

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

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

$$V_h^{\hat{\pi}} = \langle Q_h^{\hat{\pi}}, \hat{\pi}_h \rangle = \hat{\mathbb{J}}_h Q_h^{\hat{\pi}}, \quad \forall x \quad (\text{F.12})$$

$$V_{H+1}^{\hat{\pi}} = 0. \quad (\text{F.13})$$

Proof of Lemma D.1. We first write

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

Next we analyze the two terms separately.

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

$$V_h^* - \hat{V}_h = \langle \pi_h^*, Q_h^* \rangle - \langle \hat{Q}_h, \hat{\pi}_h \rangle \quad (\text{F.14})$$

$$= \mathbb{J}_h Q_h^* - \hat{\mathbb{J}}_h \hat{Q}_h \quad (\text{F.15})$$

$$= \mathbb{J}_h (Q_h^* - \hat{Q}_h) + (\mathbb{J}_h - \hat{\mathbb{J}}_h) \hat{Q}_h \quad (\text{F.16})$$

$$= \mathbb{J}_h (Q_h^* - \hat{Q}_h) + \xi_h \quad (\text{F.17})$$

$$= \mathbb{J}_h ([r_h + \mathbb{P}V_{h+1}^*] - [r_h + \mathbb{P}\hat{V}_{h+1} - \iota_h]) + \xi_h \quad (\text{F.18})$$

$$= \mathbb{J}_h \mathbb{P}(V_{h+1}^* - \hat{V}_{h+1}) + \mathbb{J}_h \iota_h + \xi_h. \quad (\text{F.19})$$

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

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

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

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

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

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

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

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

$$\hat{V}_h - V_h^{\hat{\pi}} = \hat{\mathbb{J}}_h \hat{Q}_h - \hat{\mathbb{J}}_h Q_h^{\hat{\pi}} \quad (\text{F.22})$$

$$= \hat{\mathbb{J}}_h ([r_h + \mathbb{P}\hat{V}_{h+1} - \iota_h] - [r_h + \mathbb{P}V_h^{\hat{\pi}}]) \quad (\text{F.23})$$

$$= \hat{\mathbb{J}}_h \mathbb{P}(\hat{V}_{h+1} - V_{h+1}^{\hat{\pi}}) - \hat{\mathbb{J}}_h \iota_h. \quad (\text{F.24})$$

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

Apply the recursion repeatedly we obtain

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

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

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

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

$$\widehat{V}_1(x) - V_1^{\widehat{\pi}}(x) = - \sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}[l_h(x_h, a_h) \mid x_1 = x]. \quad (\text{F.26})$$

Putting together (F.21) and (F.26) completes the proof of Lemma D.1. \square

F.4 Proof of Theorem 4.3

We define

$$\mu_A = \sigma_{\min}(\sqrt{A^\top A}), \quad L_A = \sigma_{\max}(\sqrt{A^\top A}), \quad (\text{F.27})$$

$$\mu_B = \sigma_{\min}(B), \quad L_B = \sigma_{\max}(B), \quad (\text{F.28})$$

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

$$W_{t+1} = W_t - \eta_t^\theta \cdot (K_t \psi_t) \phi_t^\top, \quad K_{t+1} = K_t + \eta_t^\omega \cdot (K_t \psi_t + x'_t - W_t \phi_t) \psi_t^\top. \quad (\text{F.29})$$

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

$$\Phi_i(\theta, \omega) := \theta^\top A^\top \omega - b_i^\top \omega - \frac{1}{2} \omega^\top B \omega, \quad (\text{F.30})$$

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.

$$P_i(\theta) = \max_{\omega} \Phi_i(\theta, \omega) = \frac{1}{2} (A\theta - b_i)^\top B^{-1} (A\theta - b_i) \quad (\text{F.31})$$

$$\widehat{\omega}_i(\theta) = \operatorname{argmax}_{\omega} \Phi_i(\theta, \omega) = B^{-1} (A\theta - b_i). \quad (\text{F.32})$$

Uniqueness of $\widehat{\omega}_i(\theta)$ is guaranteed by on the full-rankness of A and B (Assumption A.3). Define by $(\theta_i^{\text{sad}}, \omega_i^{\text{sad}})$ the saddle-point of the convex-concave function Φ_i . Then we have

$$\theta_i^{\text{sad}} = \operatorname{argmin}_{\theta} P_i(\theta), \quad \omega_i^{\text{sad}} = \widehat{\omega}_i(\theta_i^*). \quad (\text{F.33})$$

Due to the separable structure of the update (F.29), if we denote the iterates (W_t, K_t) by $W_t = [\theta_{1,t}, \dots, \theta_{d_x,t}]^\top$ and $K_t = [\omega_{1,t}, \dots, \omega_{d_x,t}]^\top$, then we can equivalently write the update as follows. For $i = 1, \dots, d_x$,

$$\begin{aligned} \theta_{i,t+1} &= \theta_{i,t} - \eta_t^\theta \widetilde{\nabla}_{\theta} \Phi_i(\theta_{i,t}, \omega_{i,t}) \\ &= \theta_{i,t} - \eta_t^\theta (\phi(x_t, a_t) \psi(z_t)^\top) \omega_{i,t} \end{aligned} \quad (\text{F.34})$$

$$\begin{aligned} \omega_{i,t+1} &= \omega_{i,t} + \eta_t^\omega \widetilde{\nabla}_{\omega} \Phi_i(\theta_{i,t}, \omega_{i,t}) \\ &= \omega_{i,t} + \eta_t^\omega (\phi(x_t, a_t)^\top \theta_{i,t} - x'_{i,t} - \psi(z_t)^\top \omega_{i,t}) \psi(z_t). \end{aligned} \quad (\text{F.35})$$

Denote by $(W^{\text{sad}}, K^{\text{sad}})$ the saddle-point of the problem (3.7). Let $(\theta^{\text{sad}}, \omega^{\text{sad}})$ be the saddle-point of Φ_i in (F.30). Since the minimax problem (3.7) is separable in the each coordinate in the primal and the dual variables, we have $\theta^{\text{sad}} = W_i^{\text{sad}}$ and $\omega^{\text{sad}} = K_i^{\text{sad}}$, for all $i = 1, \dots, d_x$, where W_i^{sad} is the i -th row of the matrix W^{sad} , and K_i^{sad} is the i -th 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

$$\nabla_\theta \Phi(\theta_t, \omega_t) = A^\top \omega_t, \quad \widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t; \xi_t) = (\phi(x_t, a_t) \psi(z_t)^\top) \omega_t \quad (\text{F.36})$$

$$\nabla_\omega \Phi(\theta_t, \omega_t) = A\theta_t - b - B\omega_t, \quad \widetilde{\nabla}_\omega \Phi(\theta_t, \omega_t; \xi_t) = (\phi(x_t, a_t)^\top \theta_t - x'_t - \psi(z_t)^\top \omega_t) \psi(z_t). \quad (\text{F.37})$$

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 (F.34) and (F.35),

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

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 $t = 1, \dots, T$. 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 $\xi_{t-1}, \theta_t, \omega_t, \widetilde{\theta}_{t+1}, \widetilde{\omega}_{t+1}$ 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 F.2. *Consider the functions P in (F.31) and Φ in (F.30).*

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 $\mu_{\text{IV}} \geq \mu_A^2/L_B$, 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 4.2) Assumptions A.4 and A.3 imply the existence and uniqueness of a matrix $W^* = [W_1^*, \dots, W_{d_x}^*] \in \mathbb{R}^{d_x \times d_\phi}$ such that $\mathbb{E}[W^* \phi(x, a) | z] = \mathbb{E}[x' | z]$. Assumption A.3 implies the uniqueness of the saddle-point $(\theta^{\text{sad}}, \omega^{\text{sad}}) = \operatorname{argmin}_{\theta \in \mathbb{R}^{d_\theta}} \max_{\omega \in \mathbb{R}^{d_\omega}} \Phi_i(\theta, \omega)$. Furthermore, in addition to Assumptions A.4 and A.3, if Assumption A.5 holds, then $W_i^* = \theta^{\text{sad}}$ and $\omega^{\text{sad}} = \widehat{\omega}_i(\widehat{\theta}) = 0$.

Proof. See §G.1. □

Item 3 above shows that under the assumptions listed in Theorem 4.3, 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

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

where $\widetilde{\nabla}_\theta \Phi$ and $\widetilde{\nabla}_\omega \Phi$ are defined in (F.36). First we show the variance of stochast gradient can be bounded by the suboptimality of the current iterate.

Lemma F.3 (Bounding variance of stochastic gradients). *Consider the sequence $\{\omega_t, \theta_t\}$. If Assumption A.2 holds, then*

$$\begin{aligned} &\mathbb{E}_t[\|\widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t) - \nabla_\theta \Phi(\theta_t, \omega_t)\|^2] \\ &\leq 4(\mu_B^{-1} \|\theta_t - \theta^*\|^2 + \|\omega_t - \widehat{\omega}(\theta_t)\|^2) + 2\sigma_{\widetilde{\nabla}_\theta}^2, \end{aligned} \quad (\text{F.38})$$

$$\begin{aligned} &\mathbb{E}_t[\|\widetilde{\nabla}_\omega \Phi(\theta_t, \omega_t) - \nabla_\omega \Phi(\theta_t, \omega_t)\|^2] \\ &\leq 16(\mu_B^{-1} \|\theta_t - \theta^*\|^2 + \|\omega_t - \widehat{\omega}(\theta_t)\|^2) + 2\sigma_{\widetilde{\nabla}_\omega}^2. \end{aligned} \quad (\text{F.39})$$

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

Proof. See §G.2. \square

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

$$\mathbb{E}[\|\theta_{t+1} - \theta^*\|^2] \leq (1 - \mu_{IV}\eta_t^\theta + 4\mu_B^{-1}(\eta_t^\theta)^2) \cdot \mathbb{E}[\|\theta_t - \theta^*\|^2] \quad (\text{F.40})$$

$$\begin{aligned} &+ (\mu_{IV}^{-1}\eta_t^\theta + 4(\eta_t^\theta)^2) \cdot \mathbb{E}[\|\omega_t - \widehat{\omega}(\theta_t)\|^2] \\ &+ 2(\eta_t^\theta)^2 \cdot \sigma_{\nabla\theta}^2 \end{aligned} \quad (\text{F.41})$$

Proof. See §G.3. \square

Lemma F.5 (One-step descent of dual update). *Consider the update sequence $\{\omega_t, \theta_t\}$. Let A.2 and A.3 hold. If $\eta_t^\omega \leq \frac{2}{\mu_B + L_B}$, then*

$$\mathbb{E}[\|\omega_{t+1} - \widehat{\omega}(\theta_{t+1})\|^2] \leq (1 - \mu_B\eta_t^\omega + 32(\mu_B^{-2}(\eta_t^\omega)^2(\eta_t^\omega)^{-1} + (\eta_t^\omega)^2 + \mu_B^{-1}(\eta_t^\omega)^2)) \cdot \mathbb{E}[\|\omega_t - \widehat{\omega}(\theta_t)\|^2]$$

$$\begin{aligned} &+ 32(\mu_B^{-2}(\eta_t^\omega)^2(\eta_t^\omega)^{-1} + \mu_B^{-1}(\eta_t^\omega)^2 + \mu_B^{-2}(\eta_t^\omega)^2) \cdot \mathbb{E}[\|\theta_t - \theta^*\|^2] \\ &+ 32((\eta_t^\omega)^2\sigma_{\nabla\omega}^2 + \mu_B^{-1}(\eta_t^\omega)^2\sigma_{\nabla\theta}^2). \end{aligned} \quad (\text{F.42})$$

Proof. See §G.4. \square

Equipped with Lemmas F.4 and F.5, we can derive a recursion by choosing appropriate stepsize sequences η_t^ω and η_t^θ . We set

$$\eta_t^\theta = \frac{\beta}{\gamma + t}, \quad \eta_t^\omega = \frac{\alpha\beta}{\gamma + t} \quad (\text{F.44})$$

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 $b_t = \mathbb{E}[\|\omega_t - \widehat{\omega}(\theta_t)\|^2]$,

$$P_t = a_t + \lambda b_t, \quad (\text{F.45})$$

and then derive a recursion formula for P_t . We have by Lemma F.4 and F.5,

$$P_{t+1} = a_{t+1} + \lambda b_{t+1} \quad (\text{F.46})$$

$$\leq (1 - \mu_{IV}\eta_t^\theta + 2^5(\lambda\alpha^{-1}\mu_B^{-2}\eta_t^\theta + \text{III}))a_t \quad (\text{F.47})$$

$$+ (1 - \mu_B\eta_t^\omega + 2^5(\alpha^{-2} \cdot \mu_B^{-2}\eta_t^\omega + \lambda^{-1}\mu_{IV}^{-1}\eta_t^\theta + \text{I}))(\lambda b_t) \quad (\text{F.48})$$

$$+ \text{II} \quad (\text{F.49})$$

where

$$\text{III} = \mu_B^{-1}(\eta_t^\theta)^2 + \lambda\mu_B^{-1}(\eta_t^\omega)^2 + \lambda\mu_B^{-2}(\eta_t^\theta)^2, \quad (\text{F.50})$$

$$\text{I} = (\eta_t^\omega)^2 + \mu_B^{-1}(\eta_t^\theta)^2 + \lambda^{-1}(\eta_t^\theta)^2, \quad (\text{F.51})$$

$$\text{II} = 2(\eta_t^\theta)^2 \cdot \sigma_{\nabla\theta}^2 + 4\lambda(\mu_B^{-1}(\eta_t^\theta)^2\sigma_{\nabla\theta}^2 + (\eta_t^\omega)^2\sigma_{\nabla\omega}^2) \quad (\text{F.52})$$

Our strategy is straight-forward. We find a suitable choice of the free parameters $(\lambda, \gamma, \alpha, \beta)$ such that 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 F.4 and F.5 are met, and (ii) the two terms 2^5III and 2^5I are less than $\frac{1}{2}\mu_{IV}\eta_t^\theta$ 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 $t \geq 1$,

$$2^5 \cdot \text{III} \leq \frac{1}{2}\mu_{IV}\eta_t^\theta \quad (\text{F.53})$$

$$2^5 \cdot \text{I} \leq \frac{1}{2}\mu_B\eta_t^\omega \quad (\text{F.54})$$

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

$$\gamma \geq 2^8 \cdot \max\{\beta \cdot \mu_B^{-1} \mu_{\text{IV}}^{-1}, \alpha^2 \lambda \beta \mu_B^{-1} \mu_{\text{IV}}^{-1}, \beta \lambda \mu_B^{-2} \mu_{\text{IV}}^{-1}, \alpha \beta \mu_B^{-1}, \alpha^{-1} \beta \mu_B^{-2}, \alpha^{-1} \lambda^{-1} \beta \mu_B^{-1}\} \quad (\text{F.55})$$

To ensure the stepsizes are small enough to meet the conditions in Lemma F.4 and F.5 we need for all t ,

$$\eta_t^\theta \leq \frac{2}{L_P + \mu_{\text{IV}}}, \quad \eta_t^\omega \leq \frac{2}{L_B + \mu_B}, \quad (\text{F.56})$$

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

$$\gamma \geq \max\{\beta, \alpha\beta\}. \quad (\text{F.57})$$

For any fixed (α, β, λ) , the inequalities (F.55) and (F.57) give the choice of γ .

Step 2. Pick α, λ such that the recursion reduces to the form $P_{t+1} \leq (1 - \frac{1}{4}\mu_{\text{IV}}\eta_t^\theta)P_t + \text{noise}$. By the choice of γ in Step 1 ((F.55) and (F.57)), the recursion (F.49) reduces to

$$P_{t+1} = a_{t+1} + \lambda b_{t+1} \quad (\text{F.58})$$

$$\leq (1 - \frac{1}{2}\mu_{\text{IV}}\eta_t^\theta + 2^5(\lambda\alpha^{-1}\mu_B^{-1}\eta_t^\theta))a_t \quad (\text{F.59})$$

$$+ (1 - \frac{1}{2}\mu_B\eta_t^\omega + 2^5(\alpha^{-2} \cdot \mu_B^{-2}\eta_t^\omega + \lambda^{-1}\mu_{\text{IV}}^{-1}\eta_t^\theta))(\lambda b_t) \quad (\text{F.60})$$

$$+ \text{II} \quad (\text{F.61})$$

We find (α, λ) such that

$$2^5(\lambda\alpha^{-1}\mu_B^{-2}\eta_t^\theta) \leq \frac{1}{4}\mu_{\text{IV}}\eta_t^\theta \quad (\text{F.62})$$

$$2^5(\alpha^{-2} \cdot \mu_B^{-2}\eta_t^\omega + \lambda^{-1}\mu_{\text{IV}}^{-1}\eta_t^\theta) \leq \frac{1}{4}\mu_B\eta_t^\omega \quad (\text{F.63})$$

It suffices to set

$$\lambda = \mu_B^{1/2} \quad (\text{F.64})$$

$$\alpha = 2^8 \cdot \mu_B^{-1.5} \mu_{\text{IV}}^{-1} \quad (\text{F.65})$$

Together the choice of λ, α in (F.64) and (F.65) implies that the recursion (F.58) simplifies to

$$P_{t+1} \leq (1 - \frac{1}{4}\mu_{\text{IV}}\eta_t^\theta)a_t + (1 - \frac{1}{4}\mu_B\eta_t^\omega)(\lambda b_t) + \text{II} \quad (\text{F.66})$$

$$\leq (1 - \frac{1}{4}\mu_{\text{IV}}\eta_t^\theta)P_t + \left(2(\eta_t^\theta)^2 \cdot \sigma_{\nabla\theta}^2 + 4\lambda(\mu_B^{-1}(\eta_t^\theta)^2\sigma_{\nabla\theta}^2 + (\eta_t^\omega)^2\sigma_{\nabla\omega}^2)\right), \quad (\text{F.67})$$

where we used $1 - \frac{1}{4}\mu_B\eta_t^\omega \leq 1 - \frac{1}{4}\mu_{\text{IV}}\eta_t^\theta$ because (F.65) implies $\alpha \geq \mu_{\text{IV}}\mu_B^{-1}$.

Next we bound the last term in (F.67). Now we study $\sigma_{\nabla\theta}^2, \sigma_{\nabla\omega}^2$. By Item 3 of Lemma F.2, we have the primal variable in the saddle-point of the minimax problem (F.30) 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 $\omega^* = 0$. Thus

$$\begin{aligned} \sigma_{\nabla\theta}^2 &= \mathbb{E}_{\xi_t} [\|\tilde{\nabla}_\theta \Phi(\theta^*, \omega^*; \xi_t)\|^2] = \mathbb{E} [\|(\phi(x_t, a_t)\psi(z_t)^\top)\omega^*\|^2] = 0 \\ \sigma_{\nabla\omega}^2 &= \mathbb{E}_{\xi_t} [\|\tilde{\nabla}_\omega \Phi(\theta^*, \omega^*; \xi_t)\|^2] \\ &= \mathbb{E} [\|(\phi(x_t, a_t)^\top \theta^* - x'_t - \psi(z_t)^\top \omega^*)\psi(z_t)\|^2] \\ &= \mathbb{E} [\|e_t\psi(z_t)\|^2] \leq \mathbb{E}[e_t^2] = \sigma^2 \end{aligned}$$

where we have used $\sup_z \|\psi(z)\|_2 \leq 1$ by A.2. This implies

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

We now restore the omitted state dimension index i , and the recursion (F.66) writes

$$\mathbb{E} [\|\theta_{t+1,i} - \theta_i^*\|^2] + \lambda \mathbb{E} [\|\omega_{t+1,i} - \hat{\omega}_i(\theta_{t+1})\|^2] \quad (\text{F.68})$$

$$\leq (1 - \frac{1}{4}\mu_{\text{IV}}\eta_t^\theta) \left(\mathbb{E} [\|\theta_{t,i} - \theta_i^*\|^2] + \lambda \mathbb{E} [\|\omega_{t,i} - \hat{\omega}_i(\theta_t)\|^2] \right) + \lambda \cdot 4(\eta_t^\omega)^2 \cdot \sigma^2. \quad (\text{F.69})$$

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

$$\tilde{P}_{t+1} \leq (1 - \frac{1}{4}\mu_{\text{IV}}\eta_t^\theta)\tilde{P}_t + \lambda \cdot 4(\eta_t^\omega)^2 \cdot d_x \sigma^2. \quad (\text{F.70})$$

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

$$\beta = 8\mu_{\text{IV}}^{-1}, \quad (\text{F.71})$$

$$\nu = \max \left\{ \gamma \tilde{P}_0, \left(\frac{1}{4}\mu_{\text{IV}}\beta - 1 \right)^{-1} \beta^2 \alpha^2 \lambda \cdot d_x \sigma^2 \right\} = \max \{ \gamma \tilde{P}_0, \text{const.} \times \mu_{\text{IV}}^{-4} \mu_B^{-2.5} \}. \quad (\text{F.72})$$

Together with our choice of α in (F.65) and λ in (F.64), we have the following choice of γ ((F.55) and (F.57))

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

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

$$\tilde{P}_t \leq \frac{\nu}{\gamma + t}. \quad (\text{F.73})$$

We prove by induction. For the base case $t = 0$, the inequality (F.73) holds by definition of ν . Next, assume for some $t \geq 0$, the inequality (F.73) holds. We investigate P_{t+1} . By the recursion formula (F.70),

$$\tilde{P}_{t+1} \leq (1 - \frac{1}{4}\mu_{\text{IV}}\eta_t^\theta)\tilde{P}_t + \lambda \cdot 4(\eta_t^\omega)^2 \cdot d_x \sigma^2 \quad (\text{F.74})$$

$$\leq \frac{\gamma + t - \frac{1}{4}\mu_{\text{IV}}\beta}{\gamma + t} \cdot \frac{\nu}{\gamma + t} + \lambda \frac{4\alpha^2 \beta^2 \cdot d_x \sigma^2}{(\gamma + t)^2} \quad (\text{F.75})$$

$$= \frac{(\gamma + t - 1)\nu}{(\gamma + t)^2} - \frac{(\frac{1}{4}\mu_{\text{IV}}\beta - 1)\nu}{(\gamma + t)^2} + \lambda \frac{4\alpha^2 \beta^2 \cdot d_x \sigma^2}{(\gamma + t)^2} \quad (\text{F.76})$$

$$\leq \frac{\nu}{\gamma + t + 1}. \quad (\text{F.77})$$

where (F.75) holds due to the recursion formula (F.70); (F.75) holds due to the induction assumption that $\tilde{P}_t \leq \nu/(\gamma + t)$; (F.77) holds because (i) $4^{-1}\mu_{\text{IV}}\beta - 1 = 1 \geq 0$ by our choice of β , and (ii) the definition of ν ensures the sum of last two terms in (F.76) is negative; (F.77) holds because $(\gamma + t - 1)/(\gamma + t)^2 \leq (\gamma + t + 1)^{-1}$. This proves the claim (F.73).

This proves Theorem 4.3.

F.5 Proof of Theorem 4.3 (ii)

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

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

Based on the error decomposition of $V^* - V^{\hat{\pi}}$ (Lemma D.1), we have for all (x, a) ,

$$\|V^* - V^{\hat{\pi}}\|_\infty = \sup_x V^*(x) - V^{\hat{\pi}}(x) \quad (\text{F.78})$$

$$\leq \sup_x \left\{ \sum_{h=1}^H \mathbb{E}_{\pi^*}[\iota_h(x_h, a_h) | x_1 = x] + \sum_{h=1}^H \mathbb{E}_{\hat{\pi}}[\iota_h(x_h, a_h) | x_1 = x] \right\}. \quad (\text{F.79})$$

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

$$\sup_{x,a} |\iota_h(x, a)| = \sup_{x,a} \left| (r_h + \mathbb{P}\widehat{V}_{h+1}) - \widehat{Q}_h \right| \quad (\text{F.80})$$

$$= \sup_{x,a} \left| (r_h + \mathbb{P}\widehat{V}_{h+1}) - (r_h + \widehat{\mathbb{P}}\widehat{V}_{h+1}) \right| \quad (\text{F.81})$$

$$= \sup_{x,a} \left| \mathbb{P}\widehat{V}_{h+1} - \widehat{\mathbb{P}}\widehat{V}_{h+1} \right| \quad (\text{F.82})$$

$$\leq \sup_{x,a} \left\{ \sqrt{\mathbb{E}_{x' \sim \mathcal{P}_{W^*}(\cdot | x, a)} [\widehat{V}_{h+1}(x')^2]} \cdot \min \left(\frac{\|(W_T - W^*)\phi(x, a)\|_2}{\sigma}, 1 \right) \right\} \quad (\text{F.83})$$

$$\leq \min \left\{ \frac{\|W_T - W^*\|}{\sigma}, 1 \right\} \cdot H. \quad (\text{F.84})$$

Here (F.81) holds by definition of \widehat{Q}_h (F.83) holds due to Lemma H.3; recall $\mathcal{P}_W(x' | x, a)$ is the probability density of multivariate Normal with mean $W\phi(x, a)$ and variance $\sigma^2 I_{d_x}$. (F.84) holds because for all $h \in [H]$ we have $\widehat{V}_h \leq H$, and that $\|(W_T - W^*)\phi(x, a)\| \leq \|W_T - W^*\| \|\phi(x, a)\|$. Note for all (x, a) we have $\|\phi(x, a)\| \leq 1$ (Assumption A.2).

Next we continue from (F.79).

$$\|V^* - V^{\widehat{\pi}}\|_\infty \leq \sup_x \left\{ \sum_{h=1}^H \mathbb{E}_{\pi^*} [\|\iota_h\|_\infty | x_1 = x] + \sum_{h=1}^H \mathbb{E}_{\widehat{\pi}} [\|\iota_h\|_\infty | x_1 = x] \right\} \quad (\text{F.85})$$

$$\leq 2H \cdot \max_{h \in [H]} \|\iota_h\|_\infty \quad (\text{F.86})$$

$$\leq 2H^2 \cdot \min \left\{ \frac{\|W_T - W^*\|}{\sigma}, 1 \right\} \leq 2H^2 \sigma^{-1} \cdot \|W_T - W^*\|. \quad (\text{F.87})$$

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

$$\mathbb{E}[\|V^* - V^{\widehat{\pi}}\|_\infty] \leq 2H^2 \sigma^{-1} \cdot \mathbb{E}[\|W_T - W^*\|] \quad (\text{F.88})$$

$$\leq 2H^2 \sigma^{-1} \cdot \sqrt{\mathbb{E}[\|W_T - W^*\|_F^2]} \leq 2H^2 \sigma^{-1} \sqrt{\frac{\nu}{\gamma + T}}. \quad (\text{F.89})$$

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

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

This completes the proof of Theorem 4.3 (ii). \square

F.6 Proof of Theorem 4.7

Proof of Theorem 4.7. Denote $\theta^{\text{sad}} = W_i^{\text{sad}}$. 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.

$$\min_{f \in \mathcal{H}_\phi} \max_{u \in \mathcal{H}_\psi} \mathbb{E}[(f(x, a) - x')u(z)] - \frac{1}{2} \mathbb{E}[u(z)^2]. \quad (\text{F.90})$$

- 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 (F.90) exists and is unique. Specifically, for $f = \theta \cdot \psi \in \mathcal{H}_\phi$, due to A.3, the inner maximization is uniquely attained at

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

Also note

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

due to the definition of the projection operator $\Pi_\psi : L^2(\mathcal{Z}) \rightarrow \mathcal{H}_\psi$, 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 $f \in \mathcal{H}_\phi$,

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

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

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

Note

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

by definition of the operator Q in Theorem 4.7. We define $\widehat{f} = \Pi_\phi f^*$, the projection of f^* onto \mathcal{H}_ϕ w.r.t the norm $\|\cdot\|_{L^2(\mathcal{S}, \mathcal{A})}$. We have the decomposition

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

For the first term we have $\|f^* - \widehat{f}\|_{L^2(\mathcal{S}, \mathcal{A})} \leq \eta_1$ by definition of η_1 . For the second term, we further decompose and use the definition of μ_{IV} and Proposition 4.1.

$$\|\widehat{f} - \theta^{\text{sad}} \cdot \phi\|_{L^2(\mathcal{S}, \mathcal{A})} \tag{F.91}$$

$$\leq \|\widehat{f} - \theta^{\text{sad}} \cdot \phi\|_\phi \tag{F.92}$$

$$\leq \mu_{\text{IV}}^{-1} \cdot \|\mathcal{T}(\widehat{f} - \theta^{\text{sad}} \cdot \phi)\|_{L^2(\mathcal{Z})} \tag{F.93}$$

$$\leq \mu_{\text{IV}}^{-1} \cdot (\|\mathcal{T}(\widehat{f} - f^*)\|_{L^2(\mathcal{Z})} + \|\mathcal{T}(f^* - \theta^{\text{sad}} \cdot \phi)\|_{L^2(\mathcal{Z})}) \tag{F.94}$$

$$\leq \mu_{\text{IV}}^{-1} \cdot (\|\mathcal{T}(\widehat{f} - f^*)\|_{L^2(\mathcal{Z})} + \|\Pi_\psi \mathcal{T}(f^* - \theta^{\text{sad}} \cdot \phi)\|_{L^2(\mathcal{Z})} + \eta_2 \cdot \mu) \tag{F.95}$$

$$\leq \mu_{\text{IV}}^{-1} \cdot (\|\mathcal{T}(\widehat{f} - f^*)\|_{L^2(\mathcal{Z})} + \|\Pi_\psi \mathcal{T}(f^* - \widehat{f})\|_{L^2(\mathcal{Z})} + \eta_2 \cdot \mu) \tag{F.96}$$

$$\leq \mu_{\text{IV}}^{-1} \cdot (2\|\mathcal{T}(\widehat{f} - f^*)\|_{L^2(\mathcal{Z})} + \eta_2 \cdot \mu) \tag{F.97}$$

$$\leq 2c \cdot \eta_1 + \mu_{\text{IV}}^{-1} \cdot \eta_2 \cdot \mu. \tag{F.98}$$

Here (F.92) follows since ϕ is bounded; (F.93) follows by definition of μ_{IV} ; (F.94) follows since \mathcal{T} is linear and we use I inequality; (F.95) follows by definition of η_2 and μ ; (F.96) follows because $\phi^\top \theta^*$ minimizes $f \mapsto \|\Pi_\psi \mathcal{T}(f^* - f)\|_{L^2(\mathcal{Z})}^2$ over \mathcal{H}_ϕ and that $\widehat{f} \in \mathcal{H}_\phi$; (F.97) follows because the projection operator is non-expansive; (F.98) follows by definition of the constant c .

This completes the proof of Theorem 4.7. \square

Remark F.6. In Theorem 4.7, the existence of such a constant c is called the stability assumption; see [9] and [19] for a detailed discussion. Note the dual approximation error η_2 is inflated by a factor of μ_{IV}^{-1} .

G Proofs of Lemmas in §F

G.1 Proof of Lemma F.2

Proof. Proof of Item 1 in Lemma F.2. 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.3)

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 $\|B^{-1/2}w\| \geq L_B^{-1/2}$. For any $v \in \mathbb{R}^{d_\phi}$ such that $\|v\| = 1$,

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

where we have used the fact that the matrix A has full column rank (Assumption A.3, $\text{rank}(A) = d_\phi$) and thus for any $u \in \mathbb{R}^{d_\phi}$ such that $\|u\| = 1$ we have $\|Au\| \geq \mu_A$. The proof of $L_P \leq L_A^2 \mu_B^{-1}$ follows by similar reasoning. To see $L_P \leq 1$, recall $D = \mathbb{E}[\phi(x, a)\phi(x, a)^\top]$. We note

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

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

Proof of Item 2 in Lemma F.2. This is obvious by noting for any θ , $\nabla_\omega^2 \Phi(\theta, \omega) = -B$ 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.3).

Proof of Item 3 in Lemma F.2. The existence of W^* such that $\mathbb{E}[W^*\phi(x, a) | z] = \mathbb{E}[x' | x, a]$ is guaranteed by Assumption A.4. 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.3.

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.3), 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 F.2, the minimum eigenvalue of $\nabla^2 P$ is bounded away from zero due to full-rankness of A and B (Assumption A.3). Thus P has a unique minimizer.

Next, we show $W_i^* = \theta^{\text{sad}}$. A.5 implies η_2 in Theorem 4.7 is zero. A.4 implies η_1 in Theorem 4.7 is zero. So Theorem 4.7 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 $\widehat{\omega}_i(\theta^{\text{sad}}) = 0$. □

G.2 Proof of Lemma F.3

Proof of Lemma F.3. For the inequality (F.38), 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 $\xi_t \notin \mathcal{F}_t$)

$$\mathbb{E}_t[\|\widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t) - \nabla_\theta \Phi(\theta_t, \omega_t)\|^2] \leq \mathbb{E}_t[\|\widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t)\|^2] \tag{G.1}$$

$$\leq 2\mathbb{E}_t[\|\widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t) - \widetilde{\nabla}_\theta \Phi(\theta^*, \omega^*)\|^2] + 2\mathbb{E}_t[\|\widetilde{\nabla}_\theta \Phi(\theta^*, \omega^*)\|^2] \tag{G.2}$$

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.2).

$$\mathbb{E}_t[\|\widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t) - \widetilde{\nabla}_\theta \Phi(\theta^*, \omega^*)\|^2] = \mathbb{E}_t[\|\phi_t \psi_t^\top (\omega_t - \omega^*)\|^2] \leq \|\omega_t - \omega^*\|^2 \tag{G.3}$$

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

$$\|\omega_t - \omega^*\|^2 \leq 2\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 2\|\widehat{\omega}(\theta_t) - \omega^*\|^2 \tag{G.4}$$

$$= 2\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 2\|(B^{-1}A)(\theta^* - \theta_t)\|^2 \tag{G.5}$$

$$\leq 2\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 2L_P \mu_B^{-1} \cdot \|\theta^* - \theta_t\|^2 \tag{G.6}$$

$$\leq 2(\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + \mu_B^{-1} \cdot \|\theta^* - \theta_t\|^2) \tag{G.7}$$

where in (G.4) we use that $\omega^* = B^{-1}(A\theta^* - b)$ and $\widehat{\omega}(\theta_t) = B^{-1}(A\theta_t - b)$; in (G.6) we use $\|B^{-1}A\| = \|B^{-1/2}(B^{-1/2}A)\| \leq \mu_B^{-1/2}L_P^{-1/2}$; in (G.7) we use $L_P \leq 1$. This completes the proof of the first inequality.

For the second inequality (F.39) we use similar reasoning.

$$\mathbb{E}_t[\|\widetilde{\nabla}_\omega\Phi(\theta_t, \omega_t) - \nabla_\omega\Phi(\theta_t, \omega_t)\|^2] \quad (\text{G.8})$$

$$\leq \mathbb{E}_t[\|\widetilde{\nabla}_\omega\Phi(\theta_t, \omega_t)\|^2] \quad (\text{G.9})$$

$$\leq 2\mathbb{E}_t[\|\widetilde{\nabla}_\omega\Phi(\theta_t, \omega_t) - \widetilde{\nabla}_\omega\Phi(\theta^*, \omega^*)\|^2] + 2\mathbb{E}_t[\|\widetilde{\nabla}_\omega\Phi(\theta^*, \omega^*)\|^2] \quad (\text{G.10})$$

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

$$\mathbb{E}_t[\|\widetilde{\nabla}_\omega\Phi(\theta_t, \omega_t) - \widetilde{\nabla}_\omega\Phi(\theta^*, \omega^*)\|^2] = \mathbb{E}_t[\|\psi_t\phi_t^\top(\theta_t - \theta^*) + \psi_t\psi_t^\top(\omega_t - \omega^*)\|^2] \quad (\text{G.11})$$

$$\leq 2\|\theta_t - \theta^*\|^2 + 2\|\omega_t - \omega^*\|^2. \quad (\text{G.12})$$

$$\leq (2 + 4L_P\mu_B^{-1})\|\theta_t - \theta^*\|^2 + 4\|\omega_t - \widehat{\omega}(\theta_t)\|^2 \quad (\text{G.13})$$

$$\leq 2^3(\mu_B^{-1}\|\theta_t - \theta^*\|^2 + \|\omega_t - \widehat{\omega}(\theta_t)\|^2) \quad (\text{G.14})$$

where we have used A.2, and $\mu_B^{-1} \geq 1$ and $L_P \leq 1$. This proves (F.39). So we complete the proof of Lemma F.3. \square

G.3 Proof Lemma F.4

Proof of Lemma F.4. Conditioning on \mathcal{F}_t , we have

$$\mathbb{E}_t[\|\theta_{t+1} - \theta^*\|^2] = \|\mathbb{E}_t[\theta_{t+1} - \theta^*]\|^2 + \mathbb{E}_t[\|(\theta_{t+1} - \theta^*) - \mathbb{E}_t[\theta_{t+1} - \theta^*]\|^2] \quad (\text{G.15})$$

We bound the first term in (G.15).

$$\|\mathbb{E}_t[\theta_{t+1} - \theta^*]\|^2 = \|\widetilde{\theta}_{t+1} - \theta^*\|^2 \quad (\text{G.16})$$

$$\leq (\|\widehat{\theta}_{t+1} - \theta^*\| + \|\widetilde{\theta}_{t+1} - \widehat{\theta}_{t+1}\|)^2 \quad (\text{G.17})$$

$$\leq ((1 - \eta_t^\theta \mu_{\text{IV}})\|\theta_t - \theta^*\| + \|\widetilde{\theta}_{t+1} - \widehat{\theta}_{t+1}\|)^2 \quad (\text{G.18})$$

$$\leq (1 - \eta_t^\theta \mu_{\text{IV}})\|\theta_t - \theta^*\|^2 + \frac{1}{\eta_t^\theta \mu_{\text{IV}}}\|\widetilde{\theta}_{t+1} - \widehat{\theta}_{t+1}\|^2, \quad (\text{G.19})$$

Here in (G.18) we use Lemma H.2 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 F.2), and (iii) our choice of stepsize. In (G.19) we use that for any $\epsilon \in (0, 1)$, it holds $((1 - \epsilon)a + b)^2 \leq (1 - \epsilon)a^2 + \epsilon^{-1}b^2$; see Lemma H.1 for a proof.

We bound the second term in (G.19) by

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

Continuing from (G.19), we have

$$\|\mathbb{E}_t[\theta_{t+1} - \theta^*]\|^2 \leq (1 - \eta_t^\theta \mu_{\text{IV}})\|\theta_t - \theta^*\|^2 + \eta_t^\theta \cdot L_A^2 \mu_{\text{IV}}^{-1} \cdot \|\omega_t - \widehat{\omega}(\theta_t)\|^2 \quad (\text{G.20})$$

Next we bound the second term in (G.15).

$$\mathbb{E}_t[\|(\theta_{t+1} - \theta^*) - \mathbb{E}_t[\theta_{t+1} - \theta^*]\|^2] = \mathbb{E}_t[\|\theta_{t+1} - \mathbb{E}_t[\theta_{t+1}]\|^2] \quad (\text{G.21})$$

$$= \mathbb{E}_t[\|\theta_{t+1} - \widetilde{\theta}_{t+1}\|^2] \quad (\text{G.22})$$

$$= (\eta_t^\theta)^2 \cdot \mathbb{E}_t[\|\widetilde{\nabla}_\theta\Phi(\theta_t, \omega_t) - \nabla\Phi(\theta_t, \omega_t)\|^2]. \quad (\text{G.23})$$

This can be bounded by Lemma F.3. Plugging into (G.15) the bounds in (G.20) and (G.23),

$$\mathbb{E}_t[\|\theta_{t+1} - \theta^*\|^2] \leq (1 - \eta_t^\theta \mu_{\text{IV}})\|\theta_t - \theta^*\|^2 + (\eta_t^\theta)L_A^2\mu_{\text{IV}}^{-1}\|\omega_t - \widehat{\omega}(\theta_t)\|^2 \quad (\text{G.24})$$

$$+ (\eta_t^\theta)^2 \mathbb{E}_t[\|\widetilde{\nabla}_\theta\Phi(\theta_t, \omega_t) - \nabla\Phi(\theta_t, \omega_t)\|^2] \quad (\text{G.25})$$

$$\leq (1 - \eta_t^\theta \mu_{\text{IV}})\|\theta_t - \theta^*\|^2 + (\eta_t^\theta)L_A^2\mu_{\text{IV}}^{-1}\|\omega_t - \widehat{\omega}(\theta_t)\|^2 \quad (\text{G.26})$$

$$+ (\eta_t^\theta)^2 \cdot (4\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 4L_P\mu_B^{-1}\|\theta_t - \theta^*\|^2 + 2\sigma_{\nabla\theta}^2) \quad (\text{G.27})$$

where we have used Lemma F.3. Taking expectation on both sides, we get

$$\mathbb{E}[\|\theta_{t+1} - \theta^*\|^2] \leq (1 - \mu_{\text{IV}}\eta_t^\theta + 4L_P\mu_B^{-1}(\eta_t^\theta)^2) \cdot \mathbb{E}[\|\theta_t - \theta^*\|^2] \quad (\text{G.28})$$

$$\begin{aligned} &+ (L_A^2\mu_{\text{IV}}^{-1}\eta_t^\theta + 4(\eta_t^\theta)^2) \cdot \mathbb{E}[\|\omega_t - \widehat{\omega}(\theta_t)\|^2] \\ &+ 2(\eta_t^\theta)^2 \cdot \sigma_{\nabla\theta}^2 \end{aligned} \quad (\text{G.29})$$

$$\leq (1 - \mu_{\text{IV}}\eta_t^\theta + 4\mu_B^{-1}(\eta_t^\theta)^2) \cdot \mathbb{E}[\|\theta_t - \theta^*\|^2] \quad (\text{G.30})$$

$$\begin{aligned} &+ (\mu_{\text{IV}}^{-1}\eta_t^\theta + 4(\eta_t^\theta)^2) \cdot \mathbb{E}[\|\omega_t - \widehat{\omega}(\theta_t)\|^2] \\ &+ 2(\eta_t^\theta)^2 \cdot \sigma_{\nabla\theta}^2 \end{aligned} \quad (\text{G.31})$$

where we use $L_P \leq 1$ and $L_A \leq 1$. This completes the proof of Lemma F.4. \square

G.4 Proof of Lemma F.5

Proof of Lemma F.5. We first bound the one-step difference of primal updates.

Lemma G.1 (One-step difference). *Consider the update sequence $\{\omega_t, \theta_t\}$. Conditioning on \mathcal{F}_t , we have*

$$\|\mathbb{E}_t[\theta_{t+1} - \theta_t]\|^2 \leq 2(\eta_t^\theta)^2(L_P^2 \cdot \|\theta_t - \theta^*\|^2 + L_A^2 \cdot \|\omega_t - \widehat{\omega}(\theta_t)\|^2). \quad (\text{G.32})$$

Proof of Lemma G.1. We start by noting

$$\|\mathbb{E}_t[\theta_{t+1} - \theta_t]\|^2 = \|\widetilde{\theta}_{t+1} - \theta_t\|^2 = (\eta_t^\theta)^2 \cdot \|A^\top \omega_t\|^2 \quad (\text{G.33})$$

$$\leq (\eta_t^\theta)^2 \cdot (2\|A^\top \widehat{\omega}(\theta_t)\|^2 + 2\|A^\top \omega_t - A^\top \widehat{\omega}(\theta_t)\|^2). \quad (\text{G.34})$$

For the first term in (G.34), we have

$$\|A^\top \widehat{\omega}(\theta_t)\| = \|\nabla P(\theta_t)\| = \|\nabla P(\theta_t) - \nabla P(\theta^*)\| \leq L_P \|\theta_t - \theta^*\|. \quad (\text{G.35})$$

For the second term in (G.34), we have

$$\|A^\top \omega_t - A^\top \widehat{\omega}(\theta_t)\|^2 \leq L_A^2 \|\omega_t - \widehat{\omega}(\theta_t)\|^2. \quad (\text{G.36})$$

Plugging into (G.34) the bounds in (G.35) and (G.36), we complete the proof of Lemma G.1. \square

Now we prove Lemma F.5. Conditioning on \mathcal{F}_t , we have

$$\mathbb{E}_t[\|\omega_{t+1} - \widehat{\omega}(\theta_{t+1})\|^2] = \|\mathbb{E}_t[\omega_{t+1} - \widehat{\omega}(\theta_{t+1})]\|^2 \quad (\text{G.37})$$

$$+ \mathbb{E}_t\left[\|(\omega_{t+1} - \widehat{\omega}(\theta_{t+1})) - \mathbb{E}_t[\omega_{t+1} - \widehat{\omega}(\theta_{t+1})]\|^2\right]. \quad (\text{G.38})$$

Next we bound the first term in (G.37)

$$\|\mathbb{E}_t[\omega_{t+1} - \widehat{\omega}(\theta_{t+1})]\|^2 = \|\mathbb{E}_t[\omega_{t+1} - \widehat{\omega}(\theta_t)] + \mathbb{E}_t[\widehat{\omega}(\theta_t) - \widehat{\omega}(\theta_{t+1})]\|^2 \quad (\text{G.39})$$

$$\leq \left(\|\mathbb{E}_t[\omega_{t+1} - \widehat{\omega}(\theta_t)]\| + \|\mathbb{E}_t[\widehat{\omega}(\theta_t) - \widehat{\omega}(\theta_{t+1})]\|\right)^2 \quad (\text{G.40})$$

$$= \left(\|\widetilde{\omega}_{t+1} - \widehat{\omega}(\theta_t)\| + \|\mathbb{E}_t[\widehat{\omega}(\theta_t) - \widehat{\omega}(\theta_{t+1})]\|\right)^2 \quad (\text{G.41})$$

$$\leq \left((1 - \mu_B\eta_t^\omega)\|\omega_t - \widehat{\omega}(\theta_t)\| + \|\mathbb{E}_t[B^{-1}A(\theta_t - \theta_{t+1})]\|\right)^2 \quad (\text{G.42})$$

$$\leq \left((1 - \mu_B\eta_t^\omega)\|\omega_t - \widehat{\omega}(\theta_t)\| + L_A\mu_B^{-1} \cdot \|\mathbb{E}_t[\theta_t - \theta_{t+1}]\|\right)^2 \quad (\text{G.43})$$

$$\leq (1 - \mu_B\eta_t^\omega)\|\omega_t - \widehat{\omega}(\theta_t)\|^2 + L_P\mu_B^{-1} \cdot \frac{1}{\mu_B\eta_t^\omega} \cdot \|\mathbb{E}_t[\theta_t - \theta_{t+1}]\|^2. \quad (\text{G.44})$$

Here in (G.42) 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 F.2), (iii) our choice of stepsize, and (iv) $\widehat{\omega}(\theta_t)$ is the minimizer of the map $\omega \mapsto -\Phi(\theta_t, \omega)$. In (G.44) we use that for any $\epsilon \in (0, 1)$, any $a, b \in \mathbb{R}$, it holds $((1 - \epsilon)a + b)^2 \leq (1 - \epsilon)a^2 + \epsilon^{-1}b^2$.

Using Lemma G.1 we can bound the second term in (G.44) by $\|\omega_t - \widehat{\omega}(\theta_t)\|$ and $\|\theta_t - \theta^*\|$.

Now we bound the second term in (G.38).

$$\mathbb{E}_t \left[\left\| (\omega_{t+1} - \widehat{\omega}(\theta_{t+1})) - \mathbb{E}_t[\omega_{t+1} - \widehat{\omega}(\theta_{t+1})] \right\|^2 \right] \quad (\text{G.45})$$

$$\leq \mathbb{E}_t \left[2 \left\| \omega_{t+1} - \mathbb{E}_t[\omega_{t+1}] \right\|^2 + 2 \left\| \widehat{\omega}(\theta_{t+1}) - \mathbb{E}_t[\widehat{\omega}(\theta_{t+1})] \right\|^2 \right]. \quad (\text{G.46})$$

For the first term in (G.46) we have

$$\mathbb{E}_t \left[\left\| \omega_{t+1} - \mathbb{E}_t[\omega_{t+1}] \right\|^2 \right] = (\eta_t^\omega)^2 \cdot \mathbb{E}_t \left[\left\| \widetilde{\nabla}_\omega \Phi(\theta_t, \omega_t) - \nabla_\omega \Phi(\theta_t, \omega_t) \right\|^2 \right] \quad (\text{G.47})$$

$$\leq (\eta_t^\omega)^2 \cdot (16 \|\theta_t - \theta^*\|^2 + 16 \|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 2\sigma_{\nabla\omega}^2) \quad (\text{G.48})$$

where we have used Lemma F.3. For the second term in (G.46) we have

$$\mathbb{E}_t \left[\left\| \widehat{\omega}(\theta_{t+1}) - \mathbb{E}_t[\widehat{\omega}(\theta_{t+1})] \right\|^2 \right] \quad (\text{G.49})$$

$$= \mathbb{E}_t \left[\left\| B^{-1} A \theta_{t+1} - \mathbb{E}_t[B^{-1} A \theta_{t+1}] \right\|^2 \right] \quad (\text{G.50})$$

$$\leq L_P \mu_B^{-1} \cdot \mathbb{E}_t \left[\left\| \theta_{t+1} - \mathbb{E}_t[\theta_{t+1}] \right\|^2 \right] \quad (\text{G.51})$$

$$= L_P \mu_B^{-1} \cdot (\eta_t^\theta)^2 \cdot \mathbb{E}_t \left[\left\| \widetilde{\nabla}_\theta \Phi(\theta_t, \omega_t) - \nabla_\theta \Phi(\theta_t, \omega_t) \right\|^2 \right] \quad (\text{G.52})$$

$$\leq \mu_B^{-1} \cdot (\eta_t^\theta)^2 \cdot (4 \|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 4 \mu_B^{-1} \|\theta_t - \theta^*\|^2 + 2\sigma_{\nabla\theta}^2). \quad (\text{G.53})$$

where we have used Lemma F.3 in (G.53).

Continuing from (G.46) (the variance part), we obtain

$$\mathbb{E}_t \left[\left\| (\omega_{t+1} - \widehat{\omega}(\theta_{t+1})) - \mathbb{E}_t[\omega_{t+1} - \widehat{\omega}(\theta_{t+1})] \right\|^2 \right] \quad (\text{G.54})$$

$$\leq 2^5 (\mu_B^{-1} (\eta_t^\omega)^2 + \mu_B^{-2} (\eta_t^\theta)^2) \|\theta_t - \theta^*\|^2 \quad (\text{G.55})$$

$$+ 2^5 ((\eta_t^\omega)^2 + \mu_B^{-1} (\eta_t^\theta)^2) \|\omega_t - \widehat{\omega}(\theta_t)\|^2 \quad (\text{G.56})$$

$$+ 4 (\mu_B^{-1} (\eta_t^\theta)^2 \sigma_{\nabla\theta}^2 + (\eta_t^\omega)^2 \sigma_{\nabla\omega}^2) \quad (\text{G.57})$$

Putting together (G.44), (G.57) and Lemma G.1, we have

$$\mathbb{E}_t \left[\left\| \omega_{t+1} - \widehat{\omega}(\theta_{t+1}) \right\|^2 \right] \leq (1 - \mu_B \eta_t^\omega) \|\omega_t - \widehat{\omega}(\theta_t)\|^2 + 2^5 (\mu_B^{-2} (\eta_t^\theta)^2 / \eta_t^\omega + \mu_B^{-1} (\eta_t^\omega)^2 + \mu_B^{-2} (\eta_t^\theta)^2) \|\theta_t - \theta^*\|^2 \quad (\text{G.58})$$

$$+ 2^5 (\mu_B^{-2} (\eta_t^\theta)^2 / \eta_t^\omega + (\eta_t^\omega)^2 + \mu_B^{-1} (\eta_t^\theta)^2) \|\omega_t - \widehat{\omega}(\theta_t)\|^2 \quad (\text{G.59})$$

$$+ 4 (\mu_B^{-1} (\eta_t^\theta)^2 \sigma_{\nabla\theta}^2 + (\eta_t^\omega)^2 \sigma_{\nabla\omega}^2). \quad (\text{G.60})$$

This completes the proof of Lemma F.5 \square

H Supporting Lemmas

Lemma H.1. 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. \square

Lemma H.2 (One-step gradient descent for smooth and strongly-convex function). Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 [25]. \square

Lemma H.3 (Expectation Difference Under Two Gaussians, Lemma C.2 in [32]). 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

$$\mathbb{E}_{z \sim \mathcal{N}_1}[g(z)] - \mathbb{E}_{z \sim \mathcal{N}_2}[g(z)] \leq \min \left\{ \frac{\|\mu_1 - \mu_2\|}{\sigma}, 1 \right\} \sqrt{\mathbb{E}_{z \sim \mathcal{N}_1}[g(z)^2]}.$$

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. □