Optimal Treatment Regimes for Personalized Medicine and Mobile Health

Eun Jeong Oh

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Penalized Q-learning for Personalized Medicine and Mobile Health

Eun Jeong Oh

There has been increasing development in personalized interventions that are tailored to uniquely evolving health status of each patient over time. In this dissertation, we investigate two problems: (1) the construction of individualized mobile health (mHealth) application recommender system; and (2) the estimation of optimal dynamic treatment regimes (DTRs) from a multi-stage clinical trial study. The dissertation is organized as follows. In Chapter 1, we provide a brief background on personalized medicine and two motivating examples which illustrate the needs and benefits of individualized treatment policies. We then introduce reinforcement learning and various methods to obtain the optimal DTRs as well as Q-learning procedure which is a popular method in the DTR literature. In Chapter 2, we propose a partial regularization via orthogonality using the adaptive Lasso (PRO-aLasso) to estimate the optimal policy which maximizes the expected utility in the mHealth setting. We also derive the convergence rate of the expected outcome of the estimated policy to that of the true optimal policy. The PRO-aLasso estimators are shown to enjoy the same oracle properties as the adaptive Lasso. Simulations and real data application demonstrate that the PRO-aLasso yields simple, more stable policies with better results as compared to the adaptive Lasso and other competing methods. In Chapter 3, we propose a penalized A-learning with a Lasso-type

penalty for the construction of optimal DTR and derive generalization error bounds of the estimated DTR. We first examine the relationship between value and the Q-functions, and then we provide a finite sample upper bound on the difference in values between the optimal DTR and the estimated DTR. In practice, we implement a multi-stage PRO-aLasso algorithm to obtain the optimal DTR. Simulation results show advantages of the proposed methods over some existing alternatives. The proposed approach is also demonstrated with the data from a depression clinical trial study. In Chapter 4, we present future work and concluding remarks.

Table of Contents

List of Tables

Acknowledgements

My deepest gratitude goes to my advisors for guiding and supporting me throughout my entire time at Columbia University. Prof. Min Qian and Prof. Ken Cheung have constantly inspired me to become a creative thinker who can focus on the details without losing sight of the big picture implications. They have instilled in me the importance of critical thinking and methodological rigor. I owe particular gratitude to Prof. Min Qian. She has been a great mentor to me, and her kind words of encouragement have helped me succeed and accomplish my goals.

I would also like to thank my dissertation committee chair, Prof. Bin Cheng, and other committee members: Prof. Caleb Miles and Prof. Ian Kronish, for sharing brilliant comments and suggestions. I thank each of them for serving on my committee and for the time they spent reviewing this work.

I would like to express my appreciation to Prof. John L. P. (Seamus) Thompson and other members at ICAP at Columbia University. Prof. Thompson has been supportive and accessible, and he has inspired me to embrace any and all academic challenges. I would also like to thank Prof. Codruta (Cody) Chiuzan for her immense support and encouragement as well as mentoring me in collaborative research. I am very grateful for the insights and guidance that I have gained from her. Acknowledgments also go to Prof. Bruce Levin, Prof. Wei-Yann Tsai, and other faculty and colleagues whom I have collaborated with.

I am also thankful to my fellow doctoral students and the department staff, Justine Herrera, Katy Hardy, Georgia A. Andre, Luminita Hellmann, and many others in the department who have supported me throughout my academic journey.

I would also like to give special thanks to Prof. Hakbae Lee in the Department of Applied Statistics at Yonsei University. He has gone above and beyond in sharing his knowledge and experience and helped me gain confidence with my choice to pursue a PhD. I am so grateful for his unwavering support and guidance. I also want to thank Seungjun Ahn for being supportive and encouraging in all my endeavours. I owe my gratitude to all of those who supported me in numerous ways throughout my doctoral studies.

Last, but certainly not least, I want to thank my family for their unconditional love and encouragement.

Chapter 1. Introduction

1.1 Overview

Deriving effective treatment rules that are tailored to a patient's individual characteristics is one of the key goals in clinical practice and medical research. Personalized medicine, sometimes referred to as precision medicine, is based on the established principle that there is no "one-size-fits-all" treatment for many heterogeneous diseases. From this perspective, it is critical to inform clinical decisions while accommodating heterogeneity among patients' drug responses. Dynamic treatment regimes (DTRs) generalize personalized interventions that are adaptive to the uniquely evolving health status of each patient over time. DTRs are alternatively known as adaptive interventions, adaptive treatment strategies or treatment policies. These policies formalize sequential individualized treatment decisions through a sequence of decision rules that map up-to-date patient information to a recommended treatment. The sequential decision rules could be, for example, intervention type, dosage level, or delivery of treatments over time. Recent efforts have targeted the development of multi-stage strategies for managing various types of chronic conditions, including depression (Lavori et al., 2000; Murphy et al., 2007; Pineau et al., 2007), diabetes (Zhao et al., 2020), HIV infection (Ernst et al., 2006; Jiang et al., 2017), and prostate cancer (Shen et al., 2017). A high-quality optimal DTR is constructed by learning a treatment rule that maximizes an empirical mean of a desired cumulative outcome; throughout, we assume that larger outcome values are preferred.

Reinforcement learning is a primary tool used to develop DTRs, where the learning behavior is through trial-and-error interactions with a dynamic environment (Kaelbling et al., 1996; Sutton & Barto, 1998). Because reinforcement learning techniques have been shown to be ef-

fective in constructing optimal DTRs, the study has attracted increasing interest in recent years among statistical researchers. Various statistical estimating methods have been extensively proposed for obtaining the optimal regimes; e.g., Q-learning (Chakraborty et al., 2010; Laber et al., 2014; Murphy, 2005; Watkins, 1989; Watkins & Dayan, 1992) and A-learning (Murphy, 2003; Robins, 2004), where Q denotes 'quality' and A denotes 'advantage'. Both Q- and A-learning utilize a backward induction algorithm to discover the optimal DTRs. However, Q-learning models the conditional mean of the response given the treatment history and covariates, whereas A-learning directly models the contrast function to find the treatment regimes. There are other approaches for specifying the optimal DTRs, including inverse probability weighting (Robins et al., 2000), augmented value maximization (Zhang et al., 2012b, 2013), and outcome weighted learning (Zhao et al., 2012). A more extensive literature review is provided in Sections 2.1 and 3.1.

In the following, we introduce two motivating examples which illustrate the needs and benefits of individualized treatment policies. Both examples aim to alleviate participants' depression and anxiety. The first example is IntelliCare, a publicly available suite of mental health and well-being applications, which was developed to address the need for diverse behavioral strategies (Lattie et al., 2016). The second example is two small depression trials for post-acute coronary syndrome (ACS) patients, coronary psychosocial evaluation studies (COPES) randomized trial (Davidson et al., 2010) and the subsequent, comparison of depression interventions after acute coronary syndrome (CODIACS) vanguard trial (Davidson et al., 2013). Both COPES and CODIACS were designed to determine the efficacy of an intervention to relieve depression for post-ACS patients. These two examples share a very important feature in common: the measurements, such as users' app usage patterns or patients' adherence to treatments, are repeatedly recorded over time. This leads to an extraordinarily large number of variables, which poses a challenge in correctly identifying the relevant features. It is apparent that variable selection is needed to remove unimportant variables from the model in a data-driven manner and

to ultimately develop optimal policies based on the predictive model. Recent developments in variable selection include shrinkage regression methods, such as least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), adaptive Lasso (Zou, 2006), smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001, 2004), and Dantzig selector (Candes & Tao, 2007). In recent years, some efforts have been made towards the development of optimal DTRs in the penalized regression framework (Lu et al., 2013; Qian & Murphy, 2011).

In what follows, we present the motivating examples about individualized treatment policies, which are in the setting of mobile health and clinical trials.

1.2 Motivating Examples

1.2.1 IntelliCare

The development of digital health interventions, including computer-assisted therapy, smartphone apps, and wearable technologies, provides an enormous potential to improve health outcomes, accessibility, clinical effectiveness, and individualization of mental health interventions (Hollis et al., 2017). With recent advance in mobile technologies, the use of health and well-being applications is substantially increasing around the world. There are over 318,000 health-related mobile applications available to users worldwide with more than 200 health applications being added each day (IQVIA, 2017). However, the absence of suitable methods of evaluating behavioral intervention technologies has been widely acknowledged (Mohr et al., 2015). With increasing adoption and usage of mobile devices, it is crucial to create a recommender system for health applications to increase user engagement and adherence, which can ultimately translate into health benefits (Christensen et al., 2009).

IntelliCare, a publicly available suite of mental health and well-being applications, was developed to address the need for diverse behavioral strategies (Lattie et al., 2016). The IntelliCare suite comprises 12 interactive applications, and each app implements a different psychologi-

cal therapy to manage mental health and wellness. For instance, iCope allows users to send themselves some inspirational messages and reassuring statements, which are written in their own words, to reduce symptoms of depression and anxiety. Primary outcomes include the Patient Health Questionnaire-9 (PHQ-9) for depression and the Generalized Anxiety Disorder-7 (GAD-7) for anxiety. Full descriptions of the IntelliCare applications can be found in Lattie et al. (2016). All these apps are managed through a Hub application. The Hub app is designed to accommodate the user's experience with the applications with its specific goal to push recommendations of apps within the suite of IntelliCare.

There is a great potential to create comprehensive and accessible mobile health interventions for clinical use. A longer-term goal of IntelliCare is to develop a recommender system that maps user's demographics, app usage patterns, and other information to a suitable and useful recommendation tailored to each individual. The recommendations through available data acquired from the suite of apps can accommodate a user's need more effectively and enhance a user's positive feelings about the recommender systems, compared to traditional psychological approaches. Therefore, it is a promising task for IntelliCare to use accumulated information from the entire population of app users to monitor efficacy and provide evidence-based recommendations as a platform. However, in the domain of mobile applications, where the data are inherently high-dimensional, developing a recommender algorithm is challenging, especially when facing high dimension, low sample size data. Thus, correctly identifying features from a large number of variables is a key to success for the development of app recommender system. This illustrates that variable selection is necessary to build a more effective, tailored intervention for each individual.

1.2.2 COPES and CODIACS Trials

Depression is the leading cause of disability around the world (Friedrich, 2017). The proportion of the worldwide population living with depression is estimated to be 4.4%, according

to a report released by the World Health Organization (WHO), which indicates that the number of the global population that suffers from depression has increased by 18.4% between 2005 and 2015 (WHO, 2017). In particular, depressive symptoms are common in patients with acute coronary syndrome (ACS) (Ellis et al., 2005). Various studies have focused on different aspects of post-ACS depression that may be associated with cardiovascular prognosis, such as depression severity (Carney et al., 2008) and persistence of the depressive episode after the acute event (Kaptein et al., 2006).

The coronary psychosocial evaluation studies (COPES) randomized trial was designed to increase the probability of demonstrating benefit from an intervention to reduce depressive symptoms in post-ACS patients (Davidson et al., 2010). In the study, 157 post-ACS patients with elevated depressive symptoms were randomly assigned among intervention options. The COPES trial, which recruited patients from hospitals in New York and Connecticut, led investigators to implement the study in multiple clinical centers. The vanguard phase of the comparison of depression interventions after acute coronary syndrome (CODIACS) was then designed for this purpose, aiming to address whether the COPES intervention can be delivered feasibly and effectively in the 5 field centers (Columbia University, New York, NY; Washington University, St Louis, MD; University of Pennsylvania, Philadelphia, PA; Emory University, Atlanta, GA; Yale University, New Haven, CT). The CODIACS vanguard trial (Davidson et al., 2013) was considered to determine the feasibility, efficacy, and costs of a centralized, stepped, patient preference-based depression care system for ACS patients. Potential participants with ACS (n = 724) were asked to participate in an eligibility interview. 177 patients were found eligible, and 150 were enrolled and randomly allocated among interventions. 150 participants were randomized to receive 6 months of treatment care, where the medication is administered at baseline and 6 months during in-person interviews and at 2 and 4 months by telephone. Patients received a centralized problem-solving treatment (PST) therapist, medication, both, or neither as an intervention option. The primary outcomes were change in Beck Depression Inventory

(BDI) scores over 6 months.

The management of post-ACS depression requires personalized, time-adaptive interventions in order to improve patient's long-term benefits, which motivated investigators to study a "clinical reinforcement learning" procedure to discover optimal personalized therapy. This procedure seeks to tailor treatment policies to patients' inherent characteristics and adapt to time-varying factors associated with disease process in order to improve the entire decision-making process. More specifically, it obtains patient responses to different regimes and maximizes the average long term outcomes as a function of patients' clinical status and multi-stage regimes using backward recursive algorithms.

In multi-stage clinical trials, the data often include a large group of predictors, so it can be high-dimensional. In CODIACS, for instance, there are a number of variables, including demographics (e.g., gender, education, and ethnicity) and 119 baseline covariates (e.g., SF-12 physical functioning scale, affinity to serotonin). Furthermore, some of these baseline measurements are repeatedly accumulated through multiple stages; i.e., the variables recorded at baseline are also observed at the subsequent time points. Estimating DTRs becomes more and more difficult as the number of variables included in the model increases. Therefore, it is necessary to remove noisy covariates and attain a smaller set of relevant variables that are important to obtain the optimal DTR. In addition, variable selection helps improve the interpretation of the DTRs, since the treatment rules with fewer variables are easier to understand.

1.3 Reinforcement Learning

Reinforcement learning (RL), a sub-area of machine learning, is one of the powerful tools used in developing dynamic treatment regimes (DTRs), where an agent learns to optimize sequences of actions in a dynamic environment through trial-and-error interactions (Kaelbling et al., 1996; Sutton & Barto, 1998). The key elements of RL procedure involve

1. State, $\boldsymbol{S}_t$,

2. Action, $\boldsymbol{A}_t$,

3. Incremental Reward, R_t,

at the t-th decision time point, $t = 0, \ldots, T$. In the typical RL approach, the agent or controller applies an action on the system and observes the corresponding reward to learn a useful control policy or action plan.

In the clinical setting, the state, $\boldsymbol{S}_t$ refers to the vector of individual covariates, such as patients' demographics, comorbidities, and concurrent medications, and the action $\boldsymbol{A}_t$ corresponds to the treatment, such as intervention type, intensity, and dosage level. Let $\bar{\boldsymbol{S}}_t = (\boldsymbol{S}_0, \ldots, \boldsymbol{S}_t)$ and $\bar{\boldsymbol{A}}_t = (\boldsymbol{A}_0, \ldots, \boldsymbol{A}_t)$ reflect the histories of state and action, respectively. The reward, R_t, represents the immediate desirability of the action chosen by the agent, which is defined as a function $f(\cdot)$ of the histories of state, $\bar{\boldsymbol{S}}_t$, the histories of action, $\bar{\boldsymbol{A}}_t$, and the next state $\boldsymbol{S}_{t+1}$; i.e., $R_t = f(\bar{\boldsymbol{S}}_t, \bar{\boldsymbol{A}}_t, \boldsymbol{S}_{t+1})$. For the realization of R_t, we use the lower case of the corresponding random variables and random vectors; namely, $r_t = f(\bar{\boldsymbol{s}}_t, \bar{\boldsymbol{a}}_t, \boldsymbol{s}_{t+1})$.

The policy $\pi_t(\bar{\boldsymbol{s}}_t, \bar{\boldsymbol{a}}_{t-1}) = \boldsymbol{a}_t, t = 0, \ldots, T$ maps from state history $\bar{\boldsymbol{s}}_t$ and early action history $\bar{\boldsymbol{a}}_{t-1}$ to the resulting action at time t, which is $\boldsymbol{a}_t$. The goal of reinforcement learning is to find the optimal regime that maximizes the expectation of the total rewards over the time trajectories, given by $\sum_{t=0}^{T} \gamma^t r_t$. The discount rate γ $(0 < \gamma < 1)$ for each time unit accommodates the weights of immediate rewards and future rewards.

The key of the dynamic programming (Bellman, 1957) is to define the optimization problem in terms of sub-problems. Given some state $\bar{\boldsymbol{S}}_t$ at time period t, the objective is to maximize the sum from t to T. We define a function $V_t(\bar{\boldsymbol{s}}_t, \bar{\boldsymbol{a}}_{t-1})$ as the value function. The value function, or simply the value, V_t is a quantity that is useful to assess the efficacy of policy. The rewards represent the immediate utility of the action, whereas the values reflect "how good" it is in the long run. The value functions used in reinforcement learning typically satisfy the recursive

Bellman equation (Bellman, 1957). Thus,

$$V_t(\bar{\boldsymbol{S}}_t, \bar{\boldsymbol{A}}_{t-1}) = \max_{\boldsymbol{a}_t} E(R_t + \gamma V_{t+1}(\bar{\boldsymbol{S}}_{t+1}, \bar{\boldsymbol{A}}_t)|\bar{\boldsymbol{S}}_t, \bar{\boldsymbol{A}}_{t-1}).$$

We denote the optimal value function as V_t^*. Then the optimal regime π_t satisfies

$$\pi_t^*(\bar{\boldsymbol{s}}_t, \bar{\boldsymbol{a}}_{t-1}) \in \arg\max_{\boldsymbol{a}_t} E(r_t + \gamma V_{t+1}^*(\bar{\boldsymbol{S}}_{t+1}, \bar{\boldsymbol{A}}_t)|\bar{\boldsymbol{S}}_t = \bar{\boldsymbol{s}}_t, \bar{\boldsymbol{A}}_{t-1} = \bar{\boldsymbol{a}}_{t-1}).$$

In the next section, we discuss various methods to estimate the optimal DTRs, and we also introduce Q-learning, which is a popular approximate dynamic programming method, with linear models.

1.3.1 Estimation Methods for Dynamic Treatment Regimes

Personalized medicine often comes into one of the two forms (Lipkovich et al., 2017): either identification of the subgroups of patients who benefit from a given treatment or identification of the optimal treatment for a specific patient. The latter form, which is also our focus, could be further divided into direct and indirect approaches:

1. Direct estimation methods, which are also known as policy search models in the reinforcement learning literature (Ng & Jordan, 2000), specify a class of regimes, $\mathcal{D}$, and directly estimate the value (reward) or marginal mean outcome for each DTR. Direct methods finally pick the DTR that maximizes the estimated value.

2. Indirect estimation approaches use a natural approximate dynamic programming to estimate stage-specific conditional mean outcomes (e.g., Q-functions) or contrasts. Then the estimated functions are maximized to infer the optimal DTR.

The popular methods among direct approaches are inverse probability weighting (Robins et al., 2000), augmented value maximization (Zhang et al., 2012b, 2013), and outcome weighted

learning (OWL), also known as O-learning (Zhao et al., 2012). In OWL, finding the optimal DTR is formulated as a weighted binary classification with the rewards as weights. Their proposed method is flexible and robust to model misspecification, but it can be less efficient if the models can be well approximated.

Two common methods using the indirect estimation approaches are Q-learning (Chakraborty et al., 2010; Laber et al., 2014; Murphy, 2005; Watkins, 1989; Watkins & Dayan, 1992) and A-learning (Murphy, 2003; Robins, 2004), where Q denotes 'quality' and A denotes 'advantage'. Q-learning is essentially a two-step procedure, where it first fits a model of the value given the interventions, covariates, and the intervention-covariate interactions and then obtains the optimal DTR based on the estimated model. The model can be parametric, semiparametric, or even nonparametric. In A-learning, proposed by Murphy (2003), one models regret functions which measure the loss incurred by not following the optimal treatment regime at each stage. Minimizing the regret functions leads to the optimal decision rule at each stage. More discussion on the relationship between Q- and A-learning can be found in Schulte et al. (2014). Both Q- and A-learning emphasize prediction accuracy of the clinical response model instead of directly optimizing the decision rule. Therefore, they rely heavily on the correctness of postulated models at all stages.

In the next section, we provide more details of Q-learning which is particularly popular in the DTR literature.

1.3.2 Q-learning with Linear Models

Consider data from n individuals in the multi-stage decision problem with a finite number of stages (say T). For each time step $t = 1, \ldots, T$, we have the vector of individual covariates, $\boldsymbol{O}_t \in \mathcal{O}_t$, and recommended actions (treatments), $\boldsymbol{A}_t \in \mathcal{A}_t$. Then as a consequence of the action, we observe the outcome of interest, Y_t, with large values desired. We assume that $\boldsymbol{A}_t$ is a categorical variable (i.e., discrete interventions). The overall outcome of interest is the sum

of stage-specific outcomes; that is, $Y = \sum_{t=1}^{T} Y_t$. In some cases, only a single terminal outcome Y_T is observed; i.e., the outcomes at all previous stages are taken to be 0. Denote the history at stage t as $\boldsymbol{H}_t = (\boldsymbol{O}_1, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_{t-1}, \boldsymbol{O}_t)$. A dynamic treatment regime (DTR), $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$, is a sequence of decision rules, where π_t is a map from the space of history, $\mathcal{H}_t$, to the action space, $\mathcal{A}_t$.

Q-learning is an approximate dynamic programming algorithm that relies on regression models for the Q-functions, where it can be viewed as an extension of least squares regression to multi-stage decision problems (Murphy, 2005). Q-learning uses a backward induction (Bellman, 1957) to first optimize the last stage treatment, then sequentially optimize the treatment in each previous stage. Dynamic programming shows that the true optimal DTR at the final stage T is $\pi_T^o(\boldsymbol{H}_T) = \arg\max_{\boldsymbol{a}_T} Q_T^o(\boldsymbol{H}_T, \boldsymbol{a}_T)$, where $Q_T^o(\boldsymbol{H}_T, \boldsymbol{A}_T) = E(Y_T|\boldsymbol{H}_T, \boldsymbol{A}_T)$, and recursively for $t = T-1, \ldots, 1$, the optimal rules are $\pi_t^o(\boldsymbol{H}_t) = \arg\max_{\boldsymbol{a}_t} Q_t^o(\boldsymbol{H}_t, \boldsymbol{a}_t)$, where $Q_t^o(\boldsymbol{H}_t, \boldsymbol{A}_t) = E(Y_t + \max_{\boldsymbol{a}_{t+1}} Q_{t+1}(\boldsymbol{H}_{t+1}, \boldsymbol{a}_{t+1})|\boldsymbol{H}_t, \boldsymbol{A}_t)$. Here, Q_t^o is called the optimal stage-t Q-function.

For simplicity, we consider the Q-learning for studies with two stages. We assume that the data with two possible treatments at each stage, $\boldsymbol{A}_t \in \{-1, 1\}$, where the treatments are randomized with known probabilities. In a two-stage study, longitudinal data on a single subject are given by the trajectory $(\boldsymbol{O}_1, \boldsymbol{A}_1, Y_1, \boldsymbol{O}_2, \boldsymbol{A}_2, Y_2)$. The histories at each stage are given by $\boldsymbol{H}_1 = \boldsymbol{O}_1$ and $\boldsymbol{H}_2 = (\boldsymbol{O}_1, \boldsymbol{A}_1, \boldsymbol{O}_2)$. The study can have either a single terminal outcome, Y, observed at the end of stage 2, or two outcomes (intermediate and final outcome), Y_1 and Y_2, observed at the end of each stage. The case of a single terminal outcome Y is viewed as a special case with $Y_1 \equiv 0$ and $Y_2 = Y$. In order to obtain a two-stage optimal DTR, say $\boldsymbol{\pi}^o = (\pi_1^o, \pi_2^o)$, we need to define the optimal Q-functions (Murphy, 2005) as follows:

$$
\begin{aligned}
Q_2^o(\boldsymbol{H}_2, \boldsymbol{A}_2) &= E(Y_2|\boldsymbol{H}_2, \boldsymbol{A}_2), \\
Q_1^o(\boldsymbol{H}_1, \boldsymbol{A}_1) &= E(Y_1 + \max_{\boldsymbol{a}_2} Q_2^o(\boldsymbol{H}_2, \boldsymbol{a}_2)|\boldsymbol{H}_1, \boldsymbol{A}_1).
\end{aligned}
$$

By using backward induction, the optimal DTR, $\boldsymbol{\pi}^o = (\pi_1^o, \pi_2^o)$, can be obtained by

$$\pi_t^o(\boldsymbol{H}_t) = \arg\max_{\boldsymbol{a}_t} Q_t^o(\boldsymbol{H}_t, \boldsymbol{a}_t), \quad t = 1, 2.$$

In practice, the true Q-functions are unknown and should be estimated from the data. Note that the Q-functions are conditional expectation and can be approximated by linear regressions. Let the Q-function for $t = 1, 2$ be modeled as

$$Q_t(\boldsymbol{H}_t, \boldsymbol{A}_t; \boldsymbol{\beta}_t, \boldsymbol{\psi}_t) = \boldsymbol{H}_{t1}^T \boldsymbol{\beta}_t + (\boldsymbol{A}_t \boldsymbol{H}_{t2})^T \boldsymbol{\psi}_t, \quad t = 1, 2, \tag{1.1}$$

where $\boldsymbol{H}_{t1}$ and $\boldsymbol{H}_{t2}$ are two known feature vectors of $\boldsymbol{H}_t$, and $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t^T, \boldsymbol{\psi}_t^T)^T$ is parameters of the Q-functions. Note that $\boldsymbol{\beta}_t$ reflects the main effect of current history on outcome, while $\boldsymbol{\psi}_t$ reflects the interaction effect of current history and treatment, which allows the treatments tailored (i.e., personalized) to each patient.

Let E_n denote the empirical expectation. The two-stage Q-learning procedures using least squares are summarized below:

1. Estimate the second-stage parameters:
   $(\hat{\boldsymbol{\beta}}_{2n}, \hat{\boldsymbol{\psi}}_{2n}) = \arg\min_{\beta_2} E_n \left(Y_2 - Q_2(\boldsymbol{H}_2, \boldsymbol{A}_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2)\right)^2$.

2. Define the first-stage pseudo-outcome:
   $\tilde{Y}_1 = Y_1 + \max_{\boldsymbol{a}_2} Q_2(\boldsymbol{H}_2, \boldsymbol{a}_2; \hat{\boldsymbol{\beta}}_{2n}, \hat{\boldsymbol{\psi}}_{2n}) = Y_1 + \boldsymbol{H}_{21}^T \hat{\boldsymbol{\beta}}_{2n} + |\boldsymbol{H}_{22}^T \hat{\boldsymbol{\psi}}_{2n}|$.

3. Estimate the first-stage parameters:
   $(\hat{\boldsymbol{\beta}}_{1n}, \hat{\boldsymbol{\psi}}_{1n}) = \arg\min_{\beta_1} E_n(\tilde{Y}_1 - Q_1(\boldsymbol{H}_1, \boldsymbol{A}_1; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1))^2$.

The quantity $\tilde{Y}_1$ is a predictor of unobserved random variable, $Y_1 + \max_{\boldsymbol{a}_2} Q_2(\boldsymbol{H}_2, \boldsymbol{a}_2)$. Once the Q-functions are estimated, one can easily obtain the optimal DTR. The estimated two-stage

optimal DTR is given by $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2)$ where the optimal regime at stage t is estimated as

$$\hat{\pi}_t(\boldsymbol{H}_t) = \arg\max_{\boldsymbol{a}_t} Q_t(\boldsymbol{H}_t, \boldsymbol{a}_t; \hat{\boldsymbol{\beta}}_{tn}, \hat{\boldsymbol{\psi}}_{tn}) = \text{sgn}(\boldsymbol{H}_{t2}^T \hat{\boldsymbol{\psi}}_{tn}), \quad t = 1, 2,$$

where $\text{sgn}(x) = 1$ if $x > 0$ and -1 otherwise.

In the next section, we take an overview of variable selection methods and variable selection in the reinforcement learning framework.

1.4 Variable Selection

Consider a regression model

$$Y = X\beta + \epsilon, \tag{1.2}$$

where $X = (x_{ij})$ is the $n \times p$ matrix of regressors, $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ is the response vector, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients, and ϵ is the vector of independent and identically distributed random errors with mean 0 and variance $\sigma^2 < \infty$. Without loss of generality, we assume that the response variable is centered and the predictors are standardized. Therefore, the linear model (1.2) does not contain the intercept.

Traditional variable selection methods involve stepwise regression, best subset selection, and some criterion-based procedures on the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC), Adjusted R^2, and Mellow's C_p statistics. Recent developments in variable selection include shrinkage regression methods, such as least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), adaptive Lasso (Zou, 2006), group Lasso (Yuan & Lin, 2006), elastic net (Zou & Hastie, 2005), and many others. In the following, we describe some of these methods used for simultaneous estimation and variable selection in the linear model.

The Lasso (Tibshirani, 1996) is a popular method which simultaneously performs variable selection and parameter estimation. It shrinks the effect of estimates with a L_1 penalty, while potentially shrinking some to be identically zero. Thus, the Lasso produces sparse solutions. The formulation of the Lasso is a convex optimization problem, and thus it can be solved quickly via coordinate descent algorithms (Efron et al., 2004; Friedman et al., 2007; Wu & Lange, 2008). The irrepresentable condition for the Lasso to consistently select the true model has been studied by Zhao and Yu (2006). Zhang and Huang (2008) examined the sparsity and bias of the Lasso selection in high-dimensional settings.

Zou (2006) proposed an adaptive Lasso penalty and presented results on model selection consistency. Zou (2006) stated that the Lasso does not have the oracle property, claiming that the model selection could be inconsistent unless the design matrix satisfies a very strong condition. The adaptive Lasso was developed to address this issue. It utilizes a Lasso-type penalty function where the weights, in the penalty term, are data dependently chosen. The weights that are used to adjust a level of penalization on individual variables. Ideally, large penalties are suitable for zero coefficients (inactive covariates) and small penalties for nonzero coefficients (active covariates). Under the appropriate choice of regularization parameter, the resulting penalized estimators enjoy the oracle property, while preserving the convexity of the Lasso. Here, the oracle property implies that the estimator can correctly identify the set of active coefficients with probability converging to one, and that the estimators of the nonzero coefficients are asymptotically normal.

Zou and Hastie (2005) introduced the elastic net, which is a convex combination of L_1 and L_2 penalties, to address the following two drawbacks of the Lasso:

1. The Lasso produces sparse solutions, but it selects at most n variables. This implies that the number of selected variables is bounded by the number of samples. When $p \gg n$, there could be more than n nonzero of β_j's in the true model. Thus, the Lasso fails

to select the correct number of variables particularly when the number of predictors far exceeds the number of observations.

2. The Lasso fails to perform grouped selection. It tends to pick only one from a group of correlated features, resulting in ignoring the grouped effect. It has also been shown that the Lasso solution paths are unstable in the presence of high collinearity, even when $p \ll n$.

The L_1 penalty attempts to generate a sparse model like the Lasso method. The L_2 penalty deals with a high correlation problem; it eliminates the the number of variables that can be selected and induces a grouping effect (grouped selection), which enables to stabilize the solution paths. Thus, the elastic net, which involves a convex combination of L_1 and L_2 penalty, has several advantages over the Lasso. Nonetheless, the elastic net still lacks the oracle property as shown in Zou and Hastie (2005). To overcome this issue, Zou and Zhang (2009) proposed an adaptive elastic net and showed that the oracle property holds under the weak regularity conditions with the diverging number of predictors. Yuan and Lin (2007) provided a necessary and sufficient condition for the elastic net to be consistent in variable selection. Jia and Yu (2008) studied conditions for selection consistency of the elastic net in the case of $p \gg n$.

Fan and Li (2001) proposed a non-convex penalty function referred to as the smoothly clipped absolute deviation (SCAD), which is a smooth transition from L_1 penalty to L_0 penalty. They demonstrated that the SCAD penalty function produces the penalized estimators possessing three desirable properties: sparsity, continuity and unbiasedness. Fan and Li (2001) and Huang and Xie (2007) established the asymptotic oracle property of the SCAD-penalized least squares estimators when the number of covariates is fixed or increases with the sample sizes.

Zhang (2007) introduced a minimax concave penalty (MCP) method. Zhang (2010) proved that the estimator enjoys variable selection consistency under the sparse Riesz condition on the design matrix. It has been also showed that MCP possesses the oracle property, which implies

that with probability tending to one, MCP can select the correct model if tuning parameters satisfy certain conditions. For more details on the properties of the MCP, see Zhang (2007, 2010).

As promising alternatives to the Lasso, non-convex penalized methods, such as the SCAD and MCP, ensures asymptotically unbiased estimates. Several authors have studied the oracle property of the folded concave penalty methods (Fan & Lv, 2011; Fan et al., 2014). However, computing the concave penalized solutions is challenging due to the non-convexity. Existing algorithms use local quadratic or local linear approximation for the concave penalty functions. More discussion on the advantages and drawbacks of such non-convex formulations can be found in Soubies et al. (2017).

1.4.1 Variable Selection in Dynamic Treatment Regimes

With the fast-paced development of technologies, investigators often gather numerous information on responses and covariates from each individual, such as patient's demographics, health records, genetic information, and molecular features. More importantly, some of those measurements are accumulated repeatedly over multiple stages (i.e., decision time points), which leads to an extraordinary large number of variables. For such big data, it is important to impose sparsity in the model to identify a smaller number of relevant variables and estimate the optimal dynamic treatment regimes (DTRs) that are more interpretable and efficient. Although there is a growing amount of study on developing variable selection methods, variable selection methods used to derive optimal DTRs have been less studied.

Existing methods for variable selection aim to minimize the prediction error, whereas the goal of variable selection in the reinforcement learning literature is to correctly identify tailoring variables which are important for constructing the optimal DTRs. Some penalization methods have been adopted to discover important variables for making individualized treatment rules. Qian and Murphy (2011) developed a two-stage procedure in the framework of Q-learning,

where they used L_1 penalized least squares to estimate optimal treatment regimes. They also studied the error bound of the value function for the estimated treatment regime. Gunter et al. (2011) proposed variable selection methods for qualitative interactions, where two variable-ranking quantities were presented. Lu et al. (2013) proposed a penalized quadratic loss in the framework of A-learning and established the oracle property of the estimator, which is robust against the misspecification of the conditional mean function. However, they only studied the case when the number of covariates is fixed and the propensity score model is known as in randomized clinical trials. Song et al. (2015a) proposed a penalized outcome weighted learning (POWL) with the fixed number of predictors, where they used the SCAD penalty to the outcome weighted learning framework to simultaneously estimate the optimal decision rule and incorporate sparsity. However, all these works only consider studies with a single treatment decision. Several improvements and extensions, such as generalization to multiple stages can be explored in future studies. There are other work in direct approaches that are developed to obtain the optimal treatment rules and considered variable selection techniques (Zhou & Kosorok, 2017a, 2017b; Zhu et al., 2015). Recently, Zhang and Zhang (2017) proposed a powerful and flexible C-learning algorithm, where C stands for 'classification'. Their proposed method learns the optimal dynamic treatment regimes backward sequentially aiming to minimize a weighted misclassification error at each stage, and uses a forward selection algorithm to choose important covariates in forming the optimal regimes. The reader may refer to Sections 2.1 and 3.1 for a more extensive literature review.

Chapter 2. Individualized mHealth Application Recommender System

2.1 Introduction

With increasing utilization of mobile devices, there is a great potential for behavioral intervention technologies via mobile applications to be included in a portfolio of available resources (Kazdin & Blase, 2011), and to be a viable option for delivering psychological treatments to mental health patients who will otherwise not have access to traditional treatments (Mohr et al., 2014). With more than 165,000 mobile health applications estimated to be available (research2gudiance, 2013), it is crucial to create a recommender system for health applications, so as to increase user engagement and adherence and ultimately lead to health benefits (Christensen et al., 2009). Our goal is to develop an individualized recommender system for health applications. Specifically, we consider building a recommender system for apps in the IntelliCare ecosystem, which is a suite of health apps for users with depression and anxiety disorders (Cheung et al., 2018; Lattie et al., 2016). Briefly, IntelliCare consists of 12 apps each implementing a psychological therapy with simple interactional elements. A Hub app is used to organize the user's experience with IntelliCare, with a specific function of pushing recommendations for other IntelliCare apps. Description of the IntelliCare apps and the effectiveness of the Hub recommendation can be found in (Lattie et al., 2016). Cheung et al. (2018) showed that the Hub recommendation is effective at increasing user engagement, and the current version of Hub makes recommendations for up to 2 apps randomly at weekly intervals. However, it is conceivable that we can further improve performance of the Hub by tailoring recommendations based on each individual's past interaction with the apps and the system's recommendation history.

One way to operationalize this type of recommender system is through a policy that takes individual information as an input and returns an action (e.g., recommendation) as an output. Our goal is to construct a high quality policy that, when implemented, will maximize the value associated with the outcome of interest. Various methods have been developed to estimate this optimal policy. Gunter et al. (2011) proposed ranking techniques designed to differentiate variables that are included merely to facilitate estimation and variables involved in the decision rules. Zhang et al. (2012b, 2013) developed an approach for estimating policy using doubly robust augmented inverse probability weighted estimator over a restricted class of regimes. Zhao et al. (2015), Zhao et al. (2012), Zhang et al. (2012a), and Zhang and Zhang (2018) proposed a statistical learning procedure, which reformulates the optimal policy estimation as a weighted classification problem. There is also a vast literature on the estimation of optimal policy based on tree-based methods (Foster et al., 2011; Laber & Zhao, 2015; Lipkovich et al., 2011; Su et al., 2008). Zhang et al. (2018), Zhang et al. (2015) and Rudin and Ertekin (2018) also proposed list-based methods which are special cases of tree-based rules.

A main challenge in developing optimal policy in our example is the high-dimensionality of the covariate space and the action space. For the IntelliCare Hub, with up to 2 recommendations among 13 apps (including Hub's self-recommendation), there are 92 possible actions. Furthermore, as in most policy development, it is imperative to consider interactions between the actions and the covariates; and this will result in a very large model that is prone to overfitting and aggravate the "curse of dimensionality". To address this challenge, researchers have applied regularization and variable selection techniques to correctly select a subset of relevant variables from the huge set of candidates. Recent developments in high-dimensional variable selection approaches include shrinkage regression methods, such as least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) method (Fan & Li, 2001, 2004), elastic net (Zou & Hastie, 2005), adaptive Lasso (Zou, 2006), and nonnegative garrote (Yuan & Lin, 2007). Penalized regression methods for estimating op-

timal policy were proposed by Qian and Murphy (2011) and Lu et al. (2013).

The foregoing variable selection methods are completely data-driven. However, there are at least two reasons for including human decisions in building a predictive model. First, investigators may have strong prior evidence that certain variables contribute much information to the response, and should be kept regardless of the data. For example, Cheung et al. (2018) demonstrated significant main effects of the Hub's recommendation; that is, users receiving a recommendation of a given app will more likely engage with the app. Keeping those *k*nown main effects in the model will intuitively improve precision in variable selection and estimation. Second, even when there is no evidence that a variable has a strong effect, the variable may be included if it facilitates interpretation of a model. We emphasize that the contribution of human expert goes beyond the identification of which variables to penalize or not to penalize in the model; it is the integration of domain knowledge into the model building and validation process.

We propose to incorporate expert knowledge in a Lasso-type variable selection procedure by performing regularization only on a pre-specified partial set of variables. Specifically, we will achieve partial regularization via an orthogonalization (PRO) technique, and apply it in conjunction with the adaptive Lasso. The remainder of the paper is organized as follows. In Section 2.2, we provide a general framework for estimating an optimal policy using the adaptive Lasso (aLasso) and applying the proposed PRO technique. In Section 2.3, we present the asymptotic behavior of our estimators and the rate of convergence for the value of the estimated policy. In Section 2.4, we compare the proposed method with some existing alternatives through extensive simulation studies. In Section 2.5, we apply the PRO-aLasso to the IntelliCare data to estimate the optimal recommender algorithm after 6 weeks of use. Discussion and conclusions are presented in Section 2.6. Proofs of theorems are included in the Appendix A.

2.2 Methodology

2.2.1 Partial Regularization via Orthogonality using the Adaptive Lasso

Suppose the observed data is of the form $\{\boldsymbol{O}, \boldsymbol{A}, Y\}$ from a sample of n individuals, where $\boldsymbol{O}$ denotes the covariates, $\boldsymbol{A}$ is the assigned actions, and Y is the outcome of interest with higher values desired. For example, in the IntelliCare data, $\boldsymbol{O}$ is the baseline number of app usage (i.e., count), $\boldsymbol{A}$ is the recommended actions by the Hub, and Y is the app usage count on log scale, observed the week after the recommendation. We assume that $\boldsymbol{A}$ is a categorical variable (i.e., discrete actions). If there are more than two actions, $\boldsymbol{A}$ is coded as a vector of dummy variables. In this context, a policy π, is a mapping from the space of observations, $\mathcal{O}$, to the action space, $\mathcal{A}$. The *value* of the policy, denoted as $V(\pi)$, is the expected outcome that would be obtained if the policy were to be implemented in the population of interest. The goal is to estimate the optimal policy, π_0, that would maximize the expected outcome if implemented:

$$\pi_0 = \arg\max_{\pi} V(\pi).$$

Define the Q-function $Q(\boldsymbol{O}, \boldsymbol{A}) = E(Y|\boldsymbol{O}, \boldsymbol{A})$ so that $Q(\boldsymbol{o}, \boldsymbol{a})$ measures the quality of assigning action $\boldsymbol{A} = \boldsymbol{a}$ to an individual with $\boldsymbol{O} = \boldsymbol{o}$ (Murphy, 2005; Qian & Murphy, 2011). Then, the optimal policy is the best action for each individual; i.e., $\pi_0(\boldsymbol{O}) = \arg\max_{\boldsymbol{a}} Q(\boldsymbol{O}, \boldsymbol{a})$. We construct the optimal policy by estimating the Q-function. We assume

$$Q(\boldsymbol{O}, \boldsymbol{A}) = \Phi(\boldsymbol{O}, \boldsymbol{A})^T \boldsymbol{\gamma}_0, \tag{2.1}$$

where $\Phi(\boldsymbol{O}, \boldsymbol{A})$ is a vector summary of $(\boldsymbol{O}, \boldsymbol{A})$. It may contain linear or higher order terms of $\boldsymbol{O}$, $\boldsymbol{A}$, and their interactions; thus, it could be high-dimensional. We separate $\Phi(\boldsymbol{O}, \boldsymbol{A})$ into two parts: those need to be penalized, denoted by $\boldsymbol{X} \in \mathbb{R}^{p_1}$, and those left unpenalized, denoted by $\boldsymbol{Z} \in \mathbb{R}^{p_2}$. Usually $\boldsymbol{Z}$ is low-dimensional and only includes several key variables. For instance,

we could let $\boldsymbol{X} = (\boldsymbol{O}, \boldsymbol{OA})$ and $\boldsymbol{Z} = (1, \boldsymbol{A})$ if the main effect of action is desired to remain in the model along with the unpenalized intercept. Thus, model (2.1) can be re-written as

$$Q(\boldsymbol{O}, \boldsymbol{A}) = \boldsymbol{X}^T\boldsymbol{\beta}_0 + \boldsymbol{Z}^T\boldsymbol{\alpha}_0, \tag{2.2}$$

where $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are the vectors of true parameters.

Although the model is high-dimensional, we expect only a few components in $\boldsymbol{X}$ are active. It is well-known that the adaptive Lasso possesses the so-called oracle properties; that is the set of non-zero coefficients is correctly identified with probability converging to one, and the estimated coefficients within this set are asymptotically normal (Zou, 2006). In what follows, we describe the PRO-aLasso algorithm that imposes an adaptive Lasso penalty only on $\boldsymbol{X}$ but not on $\boldsymbol{Z}$, and will show that the oracle properties are preserved.

Let E_n denote the sample average. The PRO-aLasso aims to find $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ that minimizes the following objective function

$$L_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) = nE_n(Y - \boldsymbol{X}^T\boldsymbol{\beta} - \boldsymbol{Z}^T\boldsymbol{\alpha})^2 + \lambda_n \sum_{j=1}^{p_1} w_j|\beta_j|, \tag{2.3}$$

and the estimated policy is the action which maximizes the estimated Q-function

$$\hat{\pi}(\boldsymbol{O}) \in \arg\max_{\boldsymbol{a}\in\boldsymbol{A}} (\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_n + \boldsymbol{Z}^T\hat{\boldsymbol{\alpha}}_n).$$

Note that λ_n in (2.3) is a tuning parameter which controls the model complexity of $\boldsymbol{X}$, and $\boldsymbol{w} = (w_1, \ldots, w_{p_1})$ is a vector of weights that are used to adjust a level of penalization on individual variables. Ideally, large penalties are suitable for zero coefficients (inactive covariates) and small penalties for non-zero coefficients (active covariates). This can be achieved by defining the weight vector as $\hat{\boldsymbol{w}} = |\bar{\boldsymbol{\beta}}|^{-\delta}$ for some $\delta > 0$ with $\bar{\boldsymbol{\beta}}$ being a root-(n/p_1)-consistent estimator. That is, heavier penalties are put on the coefficients with smaller $\bar{\boldsymbol{\beta}}$ estimates and

thus smaller true parameters. In practice, we propose to set $\bar{\boldsymbol{\beta}}$ as perturbed elastic net estimates, following Zou and Zhang (2009), and 5-fold cross-validation can be used to select an optimal pair of (δ, λ_n).

The PRO-aLasso algorithm is given below. It implies that $\hat{\boldsymbol{\beta}}_n$ is the adaptive Lasso estimator obtained based on the new response vector $\tilde{Y} = Y - \boldsymbol{Z}^T\hat{\boldsymbol{\upsilon}}_n$ and the new predictor matrix $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{Z}^T\hat{\boldsymbol{\Gamma}}_n$, where $\tilde{Y}$ is the residuals of Y on a direction orthogonal to $\boldsymbol{Z}$ (or simply Y adjusted for $\boldsymbol{Z}$) and $\tilde{\boldsymbol{X}}$ is the residuals of $\boldsymbol{X}$ on a direction orthogonal to $\boldsymbol{Z}$ (or simply $\boldsymbol{X}$ adjusted for $\boldsymbol{Z}$). The adaptive Lasso estimates can be obtained using a coordinate descent algorithm with the R package `glmnet` (Friedman et al., 2010), which is integrated in the PRO-aLasso algorithm.

Algorithm 1 PRO-aLasso Algorithm

Input: data $(\boldsymbol{O}, \boldsymbol{A}, Y)$

Output: policy $\hat{\pi}$

1: Formulate $\boldsymbol{X}$ and $\boldsymbol{Z}$ as a function of $\boldsymbol{O}$ and $\boldsymbol{A}$ in order to impose an adaptive Lasso penalty on $\boldsymbol{X}$ but not on $\boldsymbol{Z}$

2: $\hat{\boldsymbol{\upsilon}}_n \leftarrow \arg\min_{\boldsymbol{\upsilon}} E_n\left(Y - \boldsymbol{Z}^T\boldsymbol{\upsilon}\right)^2$

3: **for** $j = 1, \ldots, p_1$ **do**

4: $\quad \hat{\boldsymbol{\gamma}}_{nj} \leftarrow \arg\min_{\boldsymbol{\gamma}_j} E_n\left(X_j - \boldsymbol{Z}^T\boldsymbol{\gamma}_j\right)^2$

5: **end for**

6: $\hat{\boldsymbol{\Gamma}}_n \leftarrow (\hat{\boldsymbol{\gamma}}_{n1}, \ldots, \hat{\boldsymbol{\gamma}}_{np_1})$

7: Construct $\hat{\boldsymbol{w}} = |\bar{\boldsymbol{\beta}}|^{-\delta}$ for some $\delta > 0$ with $\bar{\boldsymbol{\beta}}$ being a root-(n/p_1)-consistent estimator, which is obtained from the response $Y - \boldsymbol{Z}^T\hat{\boldsymbol{\upsilon}}_n$ and the predictor matrix $\boldsymbol{X} - \boldsymbol{Z}^T\hat{\boldsymbol{\Gamma}}_n$

8: Define $(\boldsymbol{X} - \boldsymbol{Z}^T\hat{\boldsymbol{\Gamma}}_n)^* = (\boldsymbol{X} - \boldsymbol{Z}^T\hat{\boldsymbol{\Gamma}}_n)/\hat{\boldsymbol{w}}$

9: Solve the lasso problem for all λ_n,

$$\hat{\boldsymbol{\beta}}_n^* \leftarrow \arg\min_{\boldsymbol{\beta}} nE_n\left(Y - \boldsymbol{Z}^T\hat{\boldsymbol{\upsilon}}_n - ((\boldsymbol{X} - \boldsymbol{Z}^T\hat{\boldsymbol{\Gamma}}_n)^*)^T\boldsymbol{\beta}\right)^2 + \lambda_n \sum_{j=1}^{p_1} |\beta_j|$$

10: $\hat{\boldsymbol{\beta}}_n \leftarrow \hat{\boldsymbol{\beta}}_n^*/\hat{\boldsymbol{w}}$

11: $\hat{\boldsymbol{\alpha}}_n \leftarrow \hat{\boldsymbol{\upsilon}}_n - \hat{\boldsymbol{\Gamma}}_n\hat{\boldsymbol{\beta}}_n$

12: $\hat{\pi}(\boldsymbol{O}) \in \arg\max_{\boldsymbol{a}} \hat{Q}(\boldsymbol{O}, \boldsymbol{a}) = \arg\max_{\boldsymbol{a}}(\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_n + \boldsymbol{Z}^T\hat{\boldsymbol{\alpha}}_n)$

2.3 Theoretical Results

To study the properties of the PRO-aLasso estimator, we introduce some additional notation. Let $\mathcal{J} = \{j : \beta_{0j} \neq 0, j = 1, \ldots, p_1\}$ be the true active set of variables in $\boldsymbol{X}$, and assume that $|\mathcal{J}| = r < p_1$. Denote the estimated active set of variables by $\hat{\mathcal{J}}_n = \{j : \hat{\beta}_{nj} \neq 0, j = 1, \ldots, p_1\}$. Let $\boldsymbol{\beta}_{0\mathcal{J}} = \{\beta_{0j} : j \in \mathcal{J}\}$ and $\hat{\boldsymbol{\beta}}_{n\mathcal{J}} = \{\hat{\beta}_j : j \in \mathcal{J}\}$. Denote $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ for any $\boldsymbol{\alpha} \in \mathbb{R}^{p_2}$, $\boldsymbol{\beta} \in \mathbb{R}^{p_1}$. Then $\mathcal{S} = \{1, 2, \ldots, p_2\} \cup \{s : \theta_{0s} \neq 0,\ s = p_2 + 1, \ldots, p\}$ is the true active set of variables in $(\boldsymbol{Z}, \boldsymbol{X})$, and thus $\mathcal{J}$ is always the subset of $\mathcal{S}$. Denote

$$\boldsymbol{\Sigma} = E\begin{pmatrix} \boldsymbol{Z}\boldsymbol{Z}^T & \boldsymbol{Z}\boldsymbol{X}^T \\ \boldsymbol{X}\boldsymbol{Z}^T & \boldsymbol{X}\boldsymbol{X}^T \end{pmatrix}, \quad \boldsymbol{\Sigma}_{\mathcal{S}} = E\begin{pmatrix} \boldsymbol{Z}\boldsymbol{Z}^T & \boldsymbol{Z}\boldsymbol{X}_{\mathcal{J}}^T \\ \boldsymbol{X}_{\mathcal{J}}\boldsymbol{Z}^T & \boldsymbol{X}_{\mathcal{J}}\boldsymbol{X}_{\mathcal{J}}^T \end{pmatrix},$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Sigma}_{\mathcal{S}} \in \mathbb{R}^{(p_2+r)\times(p_2+r)}$.

Theorem 1 below shows that the PRO-aLasso estimator enjoys variable selection consistency and asymptotic normality even when the number of parameters diverges.

Theorem 1. *Suppose assumptions (A1)–(A6) in the Appendix A hold. Under model (2.2), the PRO-aLasso estimator possesses the following properties:*

i) (variable selection consistency) $\lim_n P(\hat{\mathcal{J}}_n = \mathcal{J}) = 1$,

ii) (joint asymptotic normality)

$$\sqrt{n}\psi^T \boldsymbol{\Sigma}_{\mathcal{S}}^{1/2}\begin{pmatrix} \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}_{n\mathcal{J}} - \boldsymbol{\beta}_{0\mathcal{J}} \end{pmatrix} \to_d N(0, \sigma^2),$$

where ψ is a vector of norm 1.

In the Theorem below, we provide a rate of convergence for the value of the estimated policy to that of the optimal policy. Theorem 2 advocates the approach of minimizing the estimated prediction error to estimate Q_0 and maximizing $\hat{Q}$ over $\boldsymbol{a} \in \boldsymbol{A}$ to obtain an policy.

Theorem 2. *Let $p(\boldsymbol{A}|\boldsymbol{O})$ denote the conditional distribution of action assignment given $\boldsymbol{O}$ in the training data. Suppose all assumptions in Theorem 1 hold, and $p(\boldsymbol{a}|\boldsymbol{o}) \geq S^{-1}$ for a positive constant S for all $(\boldsymbol{o}, \boldsymbol{a})$ pairs. Assume that there exist some constants $C > 0$ and $\eta \geq 0$ such that*

$$P\left(\max_{\boldsymbol{a}\in\boldsymbol{A}} Q_0(\boldsymbol{O},\boldsymbol{a}) - \max_{\boldsymbol{a}\in\boldsymbol{A}/\arg\max_{\boldsymbol{a}\in\boldsymbol{A}} Q_0(\boldsymbol{O},\boldsymbol{a})} Q_0(\boldsymbol{O},\boldsymbol{a}) \leq \epsilon\right) \leq C\epsilon^{\eta} \tag{2.4}$$

for all positive ϵ. Then

$$V(\pi_0) - V(\hat{\pi}) \leq \mathrm{O}_P\left[\left(\frac{p_2 + r}{n}\right)^{(1+\eta)/(2+\eta)}\right].$$

Remark. Condition (2.4) is a "margin" type condition. It measures the difference in mean outcomes between the optimal action(s) and the best suboptimal action(s) at $\boldsymbol{O}$. For $C = 1, \eta = 0$, condition (2.4) always holds for all $\epsilon > 0$. See Qian and Murphy (2011) for discussion of this condition.

2.4 Simulation

We conduct a set of numerical studies to assess the performance of each method. The simulated data are generated from the model

$$Y = (1, \boldsymbol{O})^T \underbrace{\begin{pmatrix}\alpha_1\\ \boldsymbol{\beta}_1\end{pmatrix}}_{\text{main effects}} + (\boldsymbol{A}, \boldsymbol{O}\boldsymbol{A})^T \underbrace{\begin{pmatrix}\alpha_2\\ \boldsymbol{\beta}_2\end{pmatrix}}_{\text{trt effects}} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. In our simulation, we set $\sigma = 1$. Covariates $\boldsymbol{O} \in \mathbb{R}^q$ are generated from $N(0, \hat{\Omega}_{q\times q})$ where $\hat{\Omega}_{q\times q}$ is a sample correlation matrix of the real data. The following is the minimum, mean, and maximum of absolute value of the correlations in $\hat{\Omega}_{q\times q}$: $(0.000, 0.076, 0.933)$. Policy action $\boldsymbol{A}$ is randomly generated from $\{-1, 1\}$ with equal probability 0.5. A number of

scenarios are considered based on the generating model that differs by the number of observations n, the number of predictors $p = 2(q+1)$, the effect size (es), and the following two cases:

1. Weak dense: $\alpha_1 = 1,\ \boldsymbol{\beta}_1 = \left\{1.25_{q/2},\ 0_{q/2}\right\},\ \boldsymbol{\beta}_2 = 4.5 \cdot |es - 0.3| \cdot \boldsymbol{\beta}_1,$

$$\alpha_2 = \frac{es \cdot \sqrt{\boldsymbol{\beta}_1^T \Omega \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2^T \Omega \boldsymbol{\beta}_2 + \sigma^2}}{2};$$

2. Sparse signal: $\alpha_1 = 1,\ \boldsymbol{\beta}_1 = \{seq(.1q + .5, 1.5),\ rep(0, .9q)\},\ \boldsymbol{\beta}_2 = 4.5 \cdot |es - 0.3| \cdot \boldsymbol{\beta}_1,$

$$\alpha_2 = \frac{es \cdot \sqrt{\boldsymbol{\beta}_1^T \Omega \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2^T \Omega \boldsymbol{\beta}_2 + \sigma^2}}{2}.$$

In settings with the weak dense scenario, half of the $\boldsymbol{\beta}_1$ components are zero. However, in the sparse signal case, nine-tenths of $\boldsymbol{\beta}_1$ are zero; e.g., if $q = 10$, $\boldsymbol{\beta}_1 = (1.5, 0, 0, \ldots, 0)$, and if $q = 40$, $\boldsymbol{\beta}_1 = (4.5, 3.5, 2.5, 1.5, 0, 0, \ldots, 0)$. In both cases, $\boldsymbol{\beta}_2$ has the same structure to $\boldsymbol{\beta}_1$ with a different magnitude.

In each scenario, we generate $n = 50$ and $n = 200$ samples with $p = 82$ (i.e., $q = 40$) and es ranging over $0, 0.2, 0.5, 0.8$. The performances of the methods are assessed by the following three aspects. The first is to evaluate the variable selection performance in $\boldsymbol{\beta}_2$ using (C, IC), where C is the number of correctly identified active variables and IC is the number of zero variables incorrectly selected in the final model, since the size of $\boldsymbol{\beta}_2$ indicates the number of tailoring variables to construct the optimal policy, $\hat{\pi}(\boldsymbol{O}) = \left|\hat{\alpha}_2 + \boldsymbol{O}^T \hat{\boldsymbol{\beta}}_2\right|$. The second is to assess the value function by the estimated optimal policy using an independent test dataset with sample size of $5,000$. The third is to estimate the root-mean-squared error (RMSE), $\sqrt{E_n(Y - \boldsymbol{X}^T \hat{\boldsymbol{\beta}}_n - \boldsymbol{Z}^T \hat{\boldsymbol{\alpha}}_n)^2}$. The simulation results are summarized in Table 2.1.

In the weak dense case, as the sample size increases, all the methods tend to correctly iden-

Table 2.1: Simulation results based on $1,000$ replications. The median number of correctly identified active variables in $\boldsymbol{\beta}_2$, denoted by C, and the median number of zero variables in $\boldsymbol{\beta}_2$ incorrectly selected in the final model, denoted by IC, are recorded along with the mean absolute deviation in parentheses. The mean of values and the root-mean-squared error (RMSE) are also reported with the standard deviation in parentheses. The best results are highlighted in boldface.

(a) Weak Dense Case

		$n = 50$				$n = 200$			
es	Method	C	IC	Value	RMSE	C	IC	Value	RMSE
0	Truth	20	0	7.19		20	0	7.19	
	PRO-aLasso	7 (2.97)	1 (1.48)	5.26 (1.14)	**7.23 (1.26)**	20 (0)	1 (1.48)	**7.18 (0.01)**	**0.69 (0.16)**
	aLasso	4 (4.45)	0 (0)	4.21 (1.87)	8.14 (1.32)	20 (0)	6 (4.45)	7.17 (0.01)	0.75 (0.13)
	Ridge	20 (0)	20 (0)	**5.72 (0.34)**	9.39 (0.5)	20 (0)	20 (0)	7.14 (0.01)	1.38 (0.21)
	Forward	13 (1.48)	11 (1.48)	1.62 (1.43)	13.6 (2.41)	20 (0)	4 (1.48)	6.6 (0.38)	5.04 (1.04)
0.2	Truth	20	0	3.13		20	0	3.13	
	PRO-aLasso	3 (1.48)	1 (1.48)	2.01 (0.44)	**4.44 (0.62)**	18 (1.48)	2 (1.48)	**3.08 (0.02)**	**0.75 (0.12)**
	aLasso	1 (1.48)	0 (0)	1.64 (0.52)	5.01 (0.8)	19 (1.48)	3 (2.97)	3.07 (0.02)	0.79 (0.14)
	Ridge	20 (0)	20 (0)	**2.05 (0.46)**	6.21 (0.31)	20 (0)	20 (0)	3.06 (0.02)	1.11 (0.14)
	Forward	12 (1.48)	11 (1.48)	1.49 (0.44)	8.04 (1.77)	20 (0)	5 (2.97)	2.76 (0.18)	3.33 (0.5)
0.5	Truth	20	0	5.5		20	0	5.5	
	PRO-aLasso	5 (2.97)	1 (1.48)	4.17 (0.58)	**5.81 (0.99)**	20 (0)	2 (1.48)	**5.48 (0.01)**	**0.69 (0.13)**
	aLasso	3 (2.97)	0 (0)	3.22 (1.31)	6.65 (1.12)	20 (0)	4 (4.45)	**5.48 (0.01)**	0.72 (0.13)
	Ridge	20 (0)	20 (0)	**4.19 (0.38)**	7.72 (0.39)	20 (0)	20 (0)	5.46 (0.01)	1.17 (0.16)
	Forward	13 (1.48)	11 (1.48)	2.38 (1.11)	10.69 (2.15)	20 (0)	5 (2.97)	5.14 (0.21)	3.63 (0.76)
0.8	Truth	20	0	12		20	0	12	
	PRO-aLasso	9 (2.97)	1 (1.48)	**10 (0.83)**	**9.46 (1.39)**	20 (0)	0 (0)	**11.99 (0.01)**	0.91 (0.32)
	aLasso	6 (2.97)	0 (0)	8.65 (2.15)	11.01 (1.88)	20 (0)	3 (2.97)	**11.99 (0.01)**	**0.82 (0.24)**
	Ridge	20 (0)	20 (0)	9.41 (0.56)	14.29 (0.7)	20 (0)	20 (0)	11.95 (0.02)	1.94 (0.28)
	Forward	14 (1.48)	11 (1.48)	5.27 (2.38)	19.97 (2.98)	20 (0)	4 (1.48)	10.75 (0.44)	11.9 (1.06)

(b) Sparse Signal Case

		$n = 50$				$n = 200$			
es	Method	C	IC	Value	RMSE	C	IC	Value	RMSE
0	Truth	4	0	7.48		4	0	7.48	
	PRO-aLasso	3 (1.48)	0 (0)	**7.25 (0.48)**	**2.66 (1.33)**	4 (0)	0 (0)	**7.48 (0.00)**	**0.39 (0.11)**
	aLasso	1 (1.48)	0 (0)	4.34 (1.84)	8.78 (1.37)	4 (0)	3 (1.48)	7.28 (0.63)	1.61 (1.93)
	Ridge	4 (0)	36 (0)	4.92 (0.49)	10.26 (0.22)	4 (0)	36 (0)	7.44 (0.02)	1.65 (0.27)
	Forward	4 (0)	20 (2.97)	1.01 (0.32)	11.3 (0.71)	4 (0)	7 (2.97)	1.38 (2.43)	9.95 (0.79)
0.2	Truth	4	0	3.36		4	0	3.36	
	PRO-aLasso	2 (0)	0 (0)	**3.11 (0.32)**	**1.76 (0.71)**	4 (0)	0 (0)	**3.35 (0.01)**	**0.33 (0.08)**
	aLasso	1 (1.48)	0 (0)	2.2 (0.64)	3.75 (1.6)	4 (0)	2 (2.97)	3.19 (0.28)	0.89 (0.67)
	Ridge	4 (0)	36 (0)	1.91 (0.37)	6.63 (0.15)	4 (0)	36 (0)	3.29 (0.02)	1.25 (0.16)
	Forward	4 (0)	19.5 (2.22)	1.83 (0.11)	6.26 (0.75)	4 (0)	6 (2.97)	1.84 (0.01)	5.46 (0.26)
0.5	Truth	4	0	5.63		4	0	5.63	
	PRO-aLasso	3 (1.48)	0 (0)	**5.47 (0.21)**	**2.02 (0.96)**	4 (0)	0 (0)	**5.63 (0.00)**	**0.35 (0.09)**
	aLasso	2 (1.48)	0 (0)	4.02 (0.76)	4.91 (1.37)	4 (0)	2 (2.97)	5.29 (0.57)	1.47 (1.61)
	Ridge	4 (0)	36 (0)	3.55 (0.42)	8.18 (0.19)	4 (0)	36 (0)	5.59 (0.02)	1.32 (0.2)
	Forward	4 (0)	20 (2.97)	3.02 (0.47)	7.9 (0.93)	4 (0)	6 (2.97)	3.17 (0.14)	6.96 (0.55)
0.8	Truth	4	0	13.07		4	0	13.07	
	PRO-aLasso	4 (0)	0 (0)	**12.89 (0.21)**	**3.56 (1.44)**	4 (0)	0 (0)	**13.06 (0.00)**	**0.56 (0.3)**
	aLasso	3 (1.48)	0 (0)	8.53 (2.35)	12.26 (2.79)	4 (0)	2 (1.48)	12.3 (1.36)	3.67 (3.75)
	Ridge	4 (0)	36 (0)	8.97 (0.83)	16.23 (0.41)	4 (0)	36 (0)	13.02 (0.02)	2.44 (0.36)
	Forward	4 (0)	20 (1.48)	7.12 (0.08)	17.02 (0.69)	4 (0)	6 (2.97)	7.14 (0.15)	16.4 (0.78)

tify the active variables. However, the PRO-aLasso selects less number of true zero variables which are incorrectly set to non-zero, compared to other competing methods. It is worth noting that good variable selection results lead to the value estimates closer to the optimal value. In the sparse signal case, the PRO-aLasso outperforms its counterparts in almost all simulation settings in terms of better variable selection performance, higher value function estimate, and smaller prediction error. In particular, our proposed method produces the estimated values nearly close to the optimal value, as the sample size grows. Not surprisingly, the forward variable selection shows a lower performance than other methods since it is highly likely to miss the ultimate model by the one-at-a-time nature of adding variables. The ridge regression performs competitively in the weak dense case but not in the sparse signal case. Based on the overall comparison between the PRO-aLasso and the adaptive Lasso (aLasso), the PRO technique seems to be a better idea in both cases.

2.5 Real Data Application

In this section, we apply our proposed method to the IntelliCare data introduced earlier. The data consists of use patterns of the 13 apps (including Hub) and the recommendation records by the Hub at 1-16 weeks after first download in $2,508$ Hub users. For illustration purposes, we apply the proposed method to estimate the optimal 6-week recommendation based on the use history in the week prior to recommendations. We considered the number of meaningful app use session ("count", O) on each app. With 13 apps including the Hub, we therefore have $\boldsymbol{O} = (O_1, \ldots, O_{13})$ from each user as the baseline covariates. We recoded each of the count variables to be 3 if greater than or equal to 3 to minimize the effect of fairly large counts. Using the notation developed above, we also let $\boldsymbol{A} = (A_1, \ldots, A_{13})$ indicate the recommendation action by the Hub. The primary outcome $Y = (Y_1, \ldots, Y_{13})$ is the count on log scale, observed the week after the recommendation; precisely, we added one before taking the log transformation to handle zero counts.

We applied various regularization and variable selection methods to build the model for each individual app use. For the model for app j, we postulate

$$\begin{aligned} E(Y_j|\boldsymbol{O},\boldsymbol{A}) =& (1, A_j, A_jO_j)^T\boldsymbol{\alpha}_j + (A_1,\ldots,A_{j-1},A_{j+1},\ldots,A_{13},\\ & O_1,\ldots,O_{13},A_jO_1,\ldots,A_jO_{j-1},A_jO_{j+1},\ldots,A_jO_{13})^T\boldsymbol{\beta}_j, \end{aligned}$$

where $\boldsymbol{\alpha}_j \in \mathbb{R}^3$ and $\boldsymbol{\beta}_j \in \mathbb{R}^{37}$. This model allows for the possibility that the use history of other apps *may* have an effect on the use of app j; however, regularization will be applied to avoid overfitting when we build the prediction model and the recommendation algorithm. On the other hand, because we expect the recommendation of app j will have a direct effect on the use pattern of app j, and are interested in estimating this effect, the PRO-aLasso does not place a penalty on the intercept, A_j, and A_jO_j (which we call concordant interaction). That is, $\boldsymbol{Z} = (1, A_j, A_jO_j)$.

We also considered maximizing the total usage as an outcome, for which we postulate

$$\begin{aligned} E(\textstyle\sum_{j=1}^{13}Y_j|\boldsymbol{O},\boldsymbol{A}) =& (1, A_1,\ldots,A_{13},A_1O_1,\ldots,A_{13}O_{13})^T\boldsymbol{\alpha}\\ &+ (O_1,\ldots,O_{13},A_1O_2,A_1O_3,\ldots,A_{13}O_{12})^T\boldsymbol{\beta}, \end{aligned}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{27}$ and $\boldsymbol{\beta} \in \mathbb{R}^{169}$. Like the individual app models, the PRO-aLasso does not place a penalty on the intercept, the direct main effects $\{A_j, j = 1,\ldots,13\}$, and the concordant interactions $\{A_jO_j, j = 1,\ldots,13\}$.

The data is randomly split so that three-fourths of data is used to estimate the optimal policy, and the remaining is used to estimate the value of the policies. In addition to the PRO-aLasso, we analyzed the data using the adaptive Lasso and the ridge regression, both of which placed penalties on all coefficients. We also ran the forward variable selection by the Akaike information criterion (AIC).

To illustrate the results of the PRO-aLasso and other methods, we examine the optimal recommender algorithm in terms of improving the usage of Daily Feats (app $j = 4$). According to the analysis results of the PRO-aLasso, the optimal decision is always to push a recommendation of Daily Feats. In contrast, the adaptive Lasso shrank all coefficients to 0 (including that of A_4) and the ridge regression yielded coefficients all very close to 0. As a result, these methods were not able to provide any policy. This indicates a pragmatic reason for avoiding penalizing the main effects.

For maximizing the total usage, the PRO-aLasso would mostly recommend the combination of Boost Me and Worry Knot in about 90.5% of the users in our data, followed by Boost Me and Day to Day (1.2%) and Boost Me and My Mantra (1.0%). The forward variable selection would recommend the combination of Thought Challenger and Worry Knot in 87.2% of the users in the data, and Hub and Worry Knot in 5.1%. The ridge regression would recommend the combination of Boost Me and Worry Knot in 68.9% and Boost Me and My Mantra in 20.7% of the users. Again, the adaptive Lasso could not provide any policy for the mobile health application recommendation.

To compare the performance of the various methods, Table 2.2 reports the estimated value and the size of policy. The size of policy is equivalent to the total number of non-zero coefficients except for the intercept and that of the baseline covariates. It is called the size of policy because it indicates the number of input variables required to operate the policy. The last column of the tables reports the observed outcome in the data set, which reflects the properties of the Hub built-in recommender system. The PRO-aLasso yielded policies with the highest values for 9 apps including ties, and this was achieved with generally small policy size. The adaptive Lasso tended to over-shrink, whereas the forward variable selection tended to overfit. Table 2.2 also compares the performance of the various methods when the objective is to maximize the total usage. The PRO-aLasso produced the policy with the highest value. Using the algorithm by PRO-aLasso would increase the sum of log-transformed number of app use sessions by 4.65

in a week on average, whereas the other methods did not appear to make a significant increase. By examining the size of policy, the PRO-aLasso seems to strike the right balance between over shrinkage (cf. aLasso) and overfitting (cf. forward variable selection).

Table 2.2: Estimated value and size of policy in parentheses for maximizing the app use count on log scale. The size of policy is the total number of non-zero coefficients except the intercept and baseline covariates. Numbers associated with the highest value are in boldface.

	PRO-aLasso	aLasso	Ridge	Forward	Observed
Individual app use					
Aspire	**0.15** (2)	**0.15** (1)	**0.15** (26)	0.07 (16)	0.05
Boost Me	0.07 (2)	0.04 (4)	0.06 (26)	**0.11** (13)	0.02
Hub	**1.20** (14)	0.78 (12)	0.78 (26)	0.34 (17)	0.33
Daily Feats	**0.08** (2)	0.03 (0)	0.03 (26)	0.07 (11)	0.03
iCope	**0.05** (2)	**0.05** (1)	0.04 (26)	**0.05** (19)	**0.05**
My Mantra	0.02 (2)	**0.07** (0)	0.02 (26)	**0.07** (16)	**0.07**
Day to Day	0.06 (2)	0.06 (2)	**0.12** (26)	0.11 (19)	0.09
MoveMe	**0.11** (2)	**0.11** (1)	0.05 (26)	0.04 (12)	0.04
Purple Chill	**0.09** (3)	**0.09** (1)	**0.09** (26)	0.05 (14)	0.04
Slumber Time	**0.07** (2)	0.03 (0)	0.03 (26)	**0.07** (10)	0.03
Social Force	**0.18** (2)	**0.18** (1)	0.07 (26)	**0.18** (17)	0.04
Thought Challenger	0.08 (2)	0.07 (4)	**0.09** (26)	0.06 (14)	0.05
Worry Knot	**0.17** (2)	**0.17** (2)	**0.17** (26)	0.14 (19)	0.04
Total usage	**4.65** (31)	0.89 (0)	4.12 (182)	2.27 (108)	0.89

2.6 Discussion

In this paper, we propose a PRO-aLasso algorithm that can be used to develop a recommender system for mobile health applications. Since the PRO involves orthogonalization, which is a linear operation, the computational cost of the PRO-aLasso is only marginally higher than that of the adaptive Lasso. The PRO technique is also versatile and can be applied with other regularization methods, although we opted to use the adaptive Lasso for its oracle properties. However, it is well-known that the oracle properties provide little value in finite samples (see the series of papers by Potscher & Leeb among others).

For the purpose of illustrating the PRO technique, we have developed the PRO-aLasso algorithm to construct one-stage policy, and applied it to the IntelliCare data for a one-off recommendation at week 6 by the Hub. We chose week 6 as the decision time, because the users would have established certain app use habits so that the weekly use data (covariates) would be relatively representative of general patterns. While the choice of week 6 is pragmatic, the Hub by design gives recommendations on a regular basis. Therefore, a realistic recommender system should provide sequential decision rules that adapt over time, and ideally account for additional personal information such as demographic variables. The versatility and computational efficiency of the PRO allow for a straightforward and feasible extension to build multi-stage policy by incorporating the PRO to reinforcement methods, such as Q-learning (Sutton & Barto, 2017). As we have demonstrated that the PRO-aLasso is superior to some common existing methods in producing one-stage recommender algorithm for the simplified setting, the PRO technique is set to be a promising tool for the multi-stage setting.

Chapter 3. Generalization Error Bounds of Dynamic Treatment Regimes in Penalized A-learning

3.1 Introduction

Discovering effective treatment regimes for life-threatening diseases is one of the key goals in medical research. In many trials, a drug which works effectively for one individual may not work or may cause serious adverse reactions for another. This classical "one-size-fits-all" approach is not appropriate if responses to the drug are heterogeneous among individuals. For instance, a significant proportion of treated patients with anti-thrombotic therapy for cardiovascular diseases suffers a new thrombotic event (Marin et al., 2009), and patients with different levels of psychiatric symptoms show heterogeneity in treatment responses (Piper et al., 1995). Precision medicine seeks solutions to such challenges by determining optimal patient-tailored treatments for a given disease.

There has been increasing development in personalized interventions that are adaptive to the uniquely evolving health status of each patient over time. Dynamic treatment regimes (DTRs), also known as adaptive interventions, adaptive treatment strategies or multi-stage treatment strategies, formalize sequential individualized treatment decisions through a sequence of decision rules that map up-to-date patient information to a recommended treatment. The multi-stage strategies were developed in a variety of health related areas, such as depression (Lavori et al., 2000; Murphy et al., 2007; Pineau et al., 2007), diabetes (Zhao et al., 2020), and acute HIV infection (Ernst et al., 2006; Jiang et al., 2017). The sequential decision rules could be, for example, intervention type, dosage, or delivery of treatments over time. A high-quality optimal DTR is constructed by learning a treatment rule that, when implemented, will maximize an

empirical mean of a desired cumulative outcome.

Various statistical estimating methods have been extensively proposed for obtaining the optimal treatment rules. Gunter et al. (2011) proposed ranking techniques designed to differentiate variables that are included merely to facilitate estimation and variables involved in the decision rules. Zhang et al. (2012b, 2013) developed an approach for estimating policy using doubly robust augmented inverse probability weighted estimator over a restricted class of regimes. Zhao et al. (2015), Zhao et al. (2012), Zhang et al. (2012a), and Zhang and Zhang (2018) proposed a statistical learning procedure, which reformulates the optimal policy estimation as a weighted classification problem. Song et al. (2015a) proposed a sparse outcome weighted learning under the classification framework for variable selection. To further improve the finite sample performance of outcome weighted learning, Zhou et al. (2017) considered a residual weighted learning method which uses a model-based method to compute the weights, and Liu et al. (2018) proposed a robust augmentation to the weights. Recently, Qi and Liu (2018) proposed a method which directly learns the single-stage optimal individualized treatment rules without main effect model and weight specifications. There is also a vast literature on the estimation of optimal policy based on tree-based methods (Foster et al., 2011; Laber & Zhao, 2015; Lipkovich et al., 2011; Su et al., 2008). Zhang et al. (2018), Zhang et al. (2015) and Rudin and Ertekin (2018) also proposed list-based methods which are special cases of tree-based rules. Fan et al. (2016) proposed a sequential advantage selection method which selects important variables with a qualitative interaction in a sequential manner. Song et al. (2015b) proposed a penalized Q-learning for the optimal DTR. Their method emphasizes individual selection, which identifies individuals without treatment effects from the population; however, no theoretical justifications were provided on the value functions. Zhu et al. (2019) considered regularization in Q-learning and developed an inference for parameters in optimal dynamic treatment regimes in the presence of nonregularity. However, much less attention has been directed towards the theoretical guarantees on the value functions. Shi et al. (2018a) proposed a penalized A-learning to obtain the

optimal DTR when the number of covariates is of the non-polynomial order of the sample size. Ertefaie (2014) proposed a method to estimate the optimal DTR in infinite horizon settings, and Luckett et al. (2019) proposed an alternative method, V-learning, for infinite horizon DTR in mobile health applications. There are also other work for constructing estimated optimal treatment rules, such as maximin projection learning (Shi et al., 2018b) and quantile optimal treatment regimes (Wang et al., 2018).

Our work is motivated by the problem of constructing optimal DTRs in clinical trial studies where participants are randomly assigned multiple times in a sequential manner. We consider the development of DTRs using data from coronary psychosocial evaluation studies (COPES) trial and the comparison of depression interventions after acute coronary syndrome (CODIACS) vanguard trial. In these studies, each patient with depressive symptoms is randomized at each stage among treatment options, and the treatment decision is made at baseline, 2, 4, and 6 months; see Davidson et al. (2010) and Davidson et al. (2013) for more details. Our goal is to develop the optimal DTRs composed of a sequence of intervention decision rules that dynamically map evolving patient information to a recommended treatment. Thus, a key methodological issue is to identify features that are predictive of outcomes among a set of variables involving time-varying covariate and treatment history from patients. However, this is usually challenging in a multi-stage decision setting where a vast number of variables are accumulated as participants move through multiple stages. For instance, in CODIACS trial, at least 115 baseline covariates are observed (e.g., SF-12 physical functioning scale, affinity to serotonin), and some are repeatedly recorded over time. Such high dimensionality of data introduces a challenge in developing the optimal DTRs due to the very large search space. Furthermore, as in most treatment selection problem, it is crucial to consider interactions between treatments and covariates; and this exacerbates the problem of high dimensionality. For such big data, it is necessary to impose sparsity in the model to identify a smaller number of relevant variables and estimate the optimal DTRs that are more interpretable and efficient. There have been ex-

tensive developments in high-dimensional variable selection, such as least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001, 2004), elastic net (Zou & Hastie, 2005), and Dantzig selector (Candes & Tao, 2007) and many others. Penalized regression methods for estimating optimal treatment rules were proposed by Qian and Murphy (2011) and Lu et al. (2013). Despite the popularity of variable selection, there is scarce literature on variable selection methods for deriving optimal DTR. Furthermore, all the aforementioned work achieved high-dimensional variable selection in a completely data-driven manner.

In this paper, we propose a penalized A-learning for the construction of the optimal DTR and further propose to incorporate human decisions in variable selection procedure by performing regularization on a pre-specified partial set of variables. Following the idea of Oh et al. (2020), we adopt a partial regularization via orthogonalization using adaptive Lasso (PRO-aLasso) and extend it to the multi-stage treatment decision problem. The paper is organized as follows. In Section 3.2, we provide a general framework of obtaining optimal DTR through A-learning with the L_1 penalty. In Section 3.3, we present generalization error bounds of the estimated DTR through our proposed method. Specifically, we examine the relationship between value and the Q-functions, and then we provide a finite sample upper bound on the difference in values between the optimal DTR and the estimated DTR. In Section 3.4 and Section 3.5, we compare the PRO-aLasso method with other alternative methods through extensive simulation studies and a real data example from the clinical trial. Discussion and conclusions are presented in Section 3.6. Proofs of theorems are included in the Appendix B.

3.2 Methodology

3.2.1 Penalized A-learning for Optimal DTR

Consider a multi-stage problem with T time points (e.g., stages or decision points). Suppose we have a training set of n trajectories. For each subject, we observe a time ordered trajectory $\{O_1, A_1, O_2, A_2, \ldots, O_T, A_T, O_{T+1}\}$, where A_t is the treatment assignment at time t for $t = 1, \ldots, T$, O_1 contains baseline information, O_t is the information observed after treatment assignment at time $(t-1)$ and prior to time t for $t = 1, \ldots, T$, and O_{T+1} is information measured after the last treatment assignment. Following treatment assignment at each time point t, there is a scalar outcome Y_t. We assume that A_t takes values in a finite, discrete space $\mathcal{A}_t$, and Y_t is continuous that is coded so that higher values are preferred. The overall outcome of interest is the sum of stage-specific outcomes $Y = \sum_{t=1}^{T} Y_t$. The case of a single terminal outcome is viewed as the special case with $Y_1 = \cdots = Y_{T-1} = 0$ and $Y_T = Y$. Define the history at time t as $H_t = (O_1, A_1, O_2, A_2, \ldots, O_t)$, which takes value in space $\mathcal{H}_t$. That is, H_t contains all information available to make decision at time t. A *dynamic treatment regime* (DTR), $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$, is a sequence of decision rules, where $\pi_t : \mathcal{H}_t \to \mathcal{A}_t$ takes patient's history as input, and returns a treatment as output at time t. The *value* of DTR, denoted by $V(\boldsymbol{\pi}) \triangleq E_{\boldsymbol{\pi}}(\sum_{t=1}^{T} Y_t)$, is the expected cumulative outcome if the entire study population were to follow the regime $\boldsymbol{\pi}$ (i.e., $A_t = \pi_t(H_t)$ for $t = 1, \ldots, T$). The optimal DTR, denoted by $\boldsymbol{\pi}^o$, is the regime that when implemented will yield the maximal value, $V(\boldsymbol{\pi}^o) = \max_{\boldsymbol{\pi}} V(\boldsymbol{\pi})$.

The goal is to use the training data to estimate the optimal DTR, $\boldsymbol{\pi}^o$. We assume that $P(A_t = a_t | H_t = h_t) > 0$ for any $(h_t, a_t) \in \mathcal{H}_t \times \mathcal{A}_t$; that is, all treatments are possible for any given history at each time point t. As demonstrated in Murphy (2005), the optimal DTR is associated with optimal Q-functions via Bellman optimality equations. Specifically, define the

optimal Q-function at time T

$$Q_T^o(h_T, a_T) = E(Y_T | H_T = h_T, A_T = a_T),$$

and the optimal Q-function at time $t = T - 1, \ldots, 1$

$$Q_t^o(h_t, a_t) = E\left[Y_t + \max_{a_{t+1} \in \mathcal{A}_{t+1}} Q_{t+1}^o(H_{t+1}, a_{t+1}) \middle| H_t = h_t, A_t = a_t\right],$$

where Q stands for "quality" of the decision based on the past history. Then, using backward induction (e.g., as in dynamic programming), the optimal DTR $\boldsymbol{\pi}^o = (\pi_1^o, \ldots, \pi_T^o)$ satisfies

$$\pi_t^o(h_t) = \arg\max_{a_t \in \mathcal{A}_t} Q_t^o(h_t, a_t)$$

for $t = 1, \ldots, T$.

Quite a few methods have been proposed based on the above arguments. Q-learning is one of the most popular approaches. It aims to estimate the optimal Q-functions backwards sequentially using regression and construct the optimal DTR by choosing a treatment that maximizes the estimated Q-functions. In contrast to Q-learning, A-learning is motivated by the fact that the optimal decisions only depend on the interaction between history and treatment in the Q-functions. Murphy (2003) and Blatt et al. (2004) proposed an iterative minimization method to directly estimate the interaction part, and Robins (2004) proposed a g-estimating equation, which can be used to produce consistent estimate of the treatment by history interaction if either the main effect of history on outcome or the propensity score model is correctly specified. Details and comparison of the two versions of A-learning can be found in Moodie et al. (2007).

In this paper, we adopt the framework in Blatt et al. (2004). Note that the optimal Q-function

at each stage can be decomposed as

$$Q_t^o(H_t, A_t) = M_t^o(H_t) + U_t^o(H_t, A_t),$$

where $M_t^o(H_t) = E[Q_t^o(H_t, A_t)|H_t]$ is the main effect of H_t and $U_t^o(H_t, A_t) = Q_t^o(H_t, A_t) - E[Q_t^o(H_t, A_t)|H_t]$ is the centered treatment effect at H_t. The optimal stage-t decision only depends on U_t^o.

We propose to model $M_t^o(H_t)$ and $U_t^o(H_t, A_t)$ by $\Phi_{t1}^\top(H_t)\boldsymbol{\theta}_{t1}$ and $\Phi_{t2}^\top(H_t, A_t)\boldsymbol{\theta}_{t2}$, respectively, where $\Phi_{t1} \in \mathbb{R}^{J_{t1}}$ is a vector summary of H_t, $\Phi_{t2} \in \mathbb{R}^{J_{t2}}$ is a vector summary of (H_t, A_t), and $\boldsymbol{\theta}_{t1}$ and $\boldsymbol{\theta}_{t2}$ are the corresponding parameters. Since $E[U_t^o(H_t, A_t)|H_t] = 0$, in practical implementation, we will center $\Phi_{t2}^\top(H_t, A_t)$ by its conditional mean $E[\Phi_{t2}^\top(H_t, A_t)|H_t]$. This can be easily done in sequentially randomized trials where the propensity score is known. Otherwise, we can plug in a propensity score estimate.

Denote $\Phi_t(H_t, A_t) = (\Phi_{t1}(H_t)^\top, \Phi_{t2}(H_t, A_t)^\top)^\top$. This gives a working model for Q-function

$$Q_t(H_t, A_t; \boldsymbol{\theta}_t) = \Phi_t(H_t, A_t)^\top \boldsymbol{\theta}_t = \Phi_{t1}(H_t)^\top \boldsymbol{\theta}_{t1} + \Phi_{t2}(H_t, A_t)^\top \boldsymbol{\theta}_{t2}, \tag{3.1}$$

where $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_{t1}^\top, \boldsymbol{\theta}_{t2}^\top)^\top \in \mathbb{R}^{J_t}$ is the parameter of interest with $J_t = J_{t1} + J_{t2}$. By the definition of the optimal Q-functions, we can verify that

$$Q_t^o(h_t, a_t) = E\left\{Y_t + \sum_{s=t+1}^{T} \left[Y_s + \max_{a_s \in \mathcal{A}_s} Q_s^o(H_s, A_s) - Q_s^o(H_s, A_s)\right] \middle| H_t = h_t, A_t = a_t\right\}$$

for $t = T-1, \ldots, 1$. Thus, the estimate of $\boldsymbol{\theta}_t$ can be obtained by regressing an estimate of $Y_t + \sum_{s=t+1}^{T} \left[Y_s + \max_{a_s \in \mathcal{A}_s} Q_s^o(H_s, A_s) - Q_s^o(H_s, A_s)\right]$ against $Q_t(H_t, A_t; \boldsymbol{\theta}_t)$. To address the high-dimensionality problem, we propose to use regression with a Lasso-type penalty. The algorithm is as follows.

1. At stage T, estimate $\boldsymbol{\theta}_T$ by

$$\hat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}_T} \left\{ \mathbb{P}_n \left[Y_T - \Phi_T(H_T, A_T)^\top \boldsymbol{\theta}_T \right]^2 + \lambda_T \sum_{j=1}^{J_T} w_{Tj} |\theta_{Tj}| \right\},$$

where $\mathbb{P}_n$ denote the empirical average over n subjects, $w_{Tj} \geq 0$ is the weight for the j-th component of $\boldsymbol{\theta}_T$, and λ_T is a tuning parameter that controls model complexity.

2. For $t = T-1, \ldots, 1$,

   (a) construct the pseudo outcome

$$\tilde{Y}_t = Y_t + \sum_{s=t+1}^{T} \left[Y_s + \max_{a_s} \Phi_s^\top(H_s, a_s) \hat{\boldsymbol{\theta}}_s - \Phi_s^\top(H_s, A_s) \hat{\boldsymbol{\theta}}_s \right];$$

   (b) estimate $\boldsymbol{\theta}_t$ by

$$\hat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}_t} \left\{ \mathbb{P}_n \left[\tilde{Y}_t - \Phi_t^\top(H_t, A_t) \boldsymbol{\theta}_t \right]^2 + \lambda_t \sum_{j=1}^{J_t} w_{tj} |\theta_{tj}| \right\},$$

   where $w_{tj} \geq 0$ is the weight for the j-th component of $\boldsymbol{\theta}_t$, and λ_t is a tuning parameter.

3. The estimated DTR is $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_T)$ satisfies

$$\hat{\pi}_t(H_t) \in \arg\max_{a_t} (\Phi_t^\top \hat{\boldsymbol{\theta}}_t) = \arg\max_{a_t} (\Phi_{t2}^\top \hat{\boldsymbol{\theta}}_{t2}), \qquad t = 1, \ldots, T.$$

The weights w_{tj}'s in the above algorithm are used to adjust level of penalization on individual variables. For example, the weights can be set to zero for a pre-specified set of clinically important variables. Alternatively, the weights could be data dependent. For instance, in adaptive Lasso, the weights are set to be inverse proportional to the magnitude of the ordinary least square or elastic net estimate of the coefficients. This allows the weights to be

small for truly nonzero coefficients and large for zero coefficients. It is observed that the data dependent weights converges to a function of true parameters which are bounded; thus, the theoretical results will hold for data dependent weights as well. For the implementation, we adopt a partial regularization via orthogonality using adaptive Lasso (PRO-aLasso) proposed by Oh et al. (2020). The PRO technique performs regularization only on a set of variables to facilitate data-driven variable selection while accommodating human input on which variables should be included (i.e., unpenalized) in the model (see Appendix B for more details).

3.3 Generalization Error Bounds

In this section, we present generalization error bounds of the estimated DTR through L_1-penalized A-learning described in Section 3.2. The error bounds do not depend on correct specification of the treatment by covariate interactions in the Q-functions. First, we examine the relationship between value and the Q-functions, and then we provide a finite sample upper bound on the difference in values between the optimal DTR and the estimated DTR.

3.3.1 Relationship between Value and Q-functions

Assume there is some positive constant S such that the propensity score $p(A_t = a_t|H_t = h_t) \geq S^{-1}$ for all pairs (h_t, a_t), $t = 1, \ldots, T$. For any DTR $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ and any square integrable functions $\{Q_t(H_t, A_t) : t = 1, \ldots, T\}$ such that $\pi_t(H_t) \in \arg\max_{a_t} Q_t(H_t, a_t)$, Murphy (2005) showed that

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) \leq \sum_{t=1}^{T} 2S^{t/2} \left\{ E\left[\left(Q_t(H_t, A_t) - Q_t^o(H_t, A_t) \right)^2 \right] \right\}^{1/2}. \tag{3.2}$$

The left hand side of (3.2) is the reduction in value of the DTR $\boldsymbol{\pi}$ as compared to the optimal DTR, The right hand side measures the distance between Q_t and the optimal Q-functions.

In the Theorem below, we further improve the upper bound. First, we show that an up-

per bound with exponent larger than $1/2$ can be obtained under a low noise condition, which implicitly implies a faster rate of convergence. Second, as we have discussed previously, the optimal decision only depends on the interaction between treatment and history, our second bound involves only the model for $U^o(H_t, A_t)$ on the right hand side of the upper bound.

Lemma 1. *Suppose there exists a constant $S \geq 1$ such that $p(a_t|\bar{o}_t, \bar{a}_{t-1}) \geq S^{-1}$ for all $(\bar{o}_t, \bar{a}_t)$ pairs for $t = 1, \ldots, T$. Assume there exist some constants $C > 0$ and $\alpha \geq 0$ such that*

$$\mathbf{P}\Big(\exists \bar{a}_{t-1} \in \bar{\mathcal{A}}_{t-1} \; s.t. \; \max_{a_t \in \mathcal{A}_t} Q_t^o(\bar{O}_t, \bar{a}_{t-1}, a_t) - \max_{a_t \in \mathcal{A}_t \setminus \arg\max_{a_t} Q_t^o(\bar{O}_t, \bar{a}_{t-1}, a_t)} Q_t^o(\bar{O}_t, \bar{a}_{t-1}, a_t) \leq \epsilon_t \Big) \leq C\epsilon_t^{\alpha} \quad (3.3)$$

for all positive ϵ_t for $t = 1, \ldots, T$. Then for any dynamic treatment regime $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ and sequence of square integrable functions $\{Q_t(H_t, A_t) : t = 1, \ldots, T\}$ such that $\pi_t(H_t) \in \arg\max_{a_t} Q_t(H_t, a_t), t = 1, \ldots, T$, we have

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) \leq \sum_{t=1}^{T} C_{1,t} \left\{ E\left[Q_t(H_t, A_t) - Q_t^o(H_t, A_t) \right]^2 \right\}^{(1+\alpha)/(2+\alpha)}. \quad (3.4)$$

Furthermore, for any square integrable function $U_t(H_t, A_t)$ such that $\arg\max_{a_t} Q_t(H_t, a_t) = \arg\max_{a_t} U_t(H_t, a_t)$ for $t = 1, \ldots, T$. Then we have

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) \leq \sum_{t=1}^{T} C_{1,t} \left[E\left(U_t(H_t, A_t) - U_t^o(H_t, A_t) \right)^2 \right]^{(1+\alpha)/(2+\alpha)}, \quad (3.5)$$

where $C_{1,t} = [2^{2+3\alpha} S^{(1+\alpha)t} C]^{1/(2+\alpha)}$.

Remark. Condition (3.3) is a margin type condition, which is similar to the margin assumptions that are widely used in the classification context (Gey, 2012; Tsybakov, 2004). Condition (3.3) measures the difference in mean outcomes between the t-stage optimal action(s)

and the t-stage best suboptimal action(s) at $\bar{O}_t$. For instance, assume $Q_t^o(\bar{O}_t, \bar{A}_{t-1}, A_t) = O_1 A_t + (A_1, A_2, \ldots, A_{t-1})A_t$ where $O_1 \sim U(-t, t)$ and A_t takes values from $\{-1, 1\}$ with probability 0.5. Then it is easy to verify that this margin type condition holds for $C = 1/[2t]$ and $\alpha = 1$. Clearly, for $C = 1$, $\alpha = 0$, condition (3.3) holds for all $\epsilon_t > 0$, and (3.4) reduces to (3.2); see Qian and Murphy (2011) for more details.

3.3.2 Quality of the Estimated DTR

In this section, we provide finite sample upper bounds on the difference between the optimal value and the value obtained by our estimator in terms of the prediction errors resulting from the estimation of Q_t^o for $t = 1, \ldots, T$. These upper bounds guarantee that if Q_t^o is consistently estimated for $t = 1, \ldots, T$, the value of the estimated dynamic treatment regime (DTR) will converge to the optimal value.

Define

$$\boldsymbol{\theta}_T^* = \arg\min_{\boldsymbol{\theta}_T} E[Y_T - \Phi_T^\top(H_T, A_T)\boldsymbol{\theta}_T]^2,$$

and for $t = T-1, \ldots, 1$, define

$$\boldsymbol{\theta}_t^* = \arg\min_{\boldsymbol{\theta}_t} E\left\{Y_t + \sum_{s=t+1}^{T}\left[Y_s - \Phi_s^\top(H_s, A_s)\boldsymbol{\theta}_s^* + \max_{a_s}\Phi_s^\top(H_s, a_s)\boldsymbol{\theta}_s^*\right] - \Phi_t^\top\boldsymbol{\theta}_t\right\}^2.$$

For expositional simplicity, assume that $\boldsymbol{\theta}_t^*$ is unique for $t = 1, \ldots, T$. For any $0 \le \gamma < 1/2$, $\eta \ge 0$ and tuning parameter λ_t, the set Θ_t^* is defined by

$$\Theta_t^* = \left\{\boldsymbol{\theta}_t \in \mathbb{R}^{J_t} : \|\Phi_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)\|_\infty \le \eta \quad \text{and} \quad E[\Phi_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)]^2 \le \gamma^2\lambda_t^2\right\}.$$

For $t = 1, \ldots, T$, define the index sets

$$I_t(\boldsymbol{\theta}_t) = \{j \in \{1, \ldots, J_t\} : w_{tj} = 0 \text{ or } \theta_{tj} \ne 0\}$$

$$I_t^c(\boldsymbol{\theta}_t) = \{1, \ldots, J_t\} \setminus I_t(\boldsymbol{\theta}_t) = \{j \in \{1, \ldots, J_t\} : w_{tj} \neq 0 \text{ and } \theta_{tj} = 0\},$$

and denote pseudo weight for $j = 1, \ldots, J_t$.

$$\bar{w}_{tj} = w_{tj} + 1_{w_{tj}=0}.$$

Without loss of generality, we re-write $\Phi_t^\top = (\Phi_{t1}^\top(H_t), \Phi_{t2}^\top(H_t, A_t))$, so that the first J_{t1} terms $\Phi_{t1}^\top$ does not contain A_t, and the remaining $J_{t2} = J_t - J_{t1}$ terms $\Phi_{t2}^\top$ satisfies $E[\Phi_{t2}^\top(H_t, A_t)|H_t] = \mathbf{0}$ a.s., holds when $p(a_t|H_t)$ is known. Note that $I_{t2}(\boldsymbol{\theta}_t) = I_t(\boldsymbol{\theta}_t) \cap \{J_{t1} + 1, \ldots, J_t\}$.

We state the following assumptions.

(B1) For $t = 1, \ldots, T$, the error terms $\varepsilon_{ti} = Y_{ti} + \sum_{s=t+1}^{T} \left[Y_{si} + \max_{a_s} Q_s^o(H_{si}, a_{si}) - Q_s^o(H_{si}, A_{si})\right] - Q_t^o(H_{ti}, A_{ti})$, where $\varepsilon_{Ti} \equiv Y_{Ti} - Q_T^o(H_{Ti}, A_{Ti})$, $i = 1, \ldots, n$, are independent of (H_{ti}, A_{ti}), $i = 1, \ldots, n$ and are i.i.d. with $E(\varepsilon_{ti}) = 0$ and $E[|\varepsilon_{ti}|^l] \leq l! c^{l-2} \sigma^2/2$ for some $c, \sigma^2 > 0$ for all $l \geq 2$.

(B2) For $t = 1, \ldots, T$, there exist finite, positive constants u and η such that $\max_{j \in \{1,\ldots,J_t\}} \|\phi_{tj}\|_\infty / \bar{w}_{tj} \leq u$ and $\|Q_t^o - \Phi_t^\top \boldsymbol{\theta}_t^*\|_\infty \leq \eta$.

(B3) For $t = 1, \ldots, T$, the gram-matrix, $M_t = E[(\phi_{t1}/\bar{w}_{t1}, \ldots, \phi_{tJ_t}/\bar{w}_{tJ_t})^\top (\phi_{t1}/\bar{w}_{t1}, \ldots, \phi_{tJ_t}/\bar{w}_{tJ_t})]$ is positive definite, and the smallest eigenvalue is denoted by τ_t.

(B4) For $t = 1, \ldots, T$, $\max_{j \in \{1,\ldots,J_t\}} E[\phi_{tj}/\bar{w}_{tj}]^2 \leq b^2$ for some $b > 0$.

For $t = 1, \ldots, T$, define

$$\Theta_t = \left\{ \boldsymbol{\theta}_t \in \Theta_t^* : \max_{s \in \{t,\ldots,T\}} \{|I_s(\boldsymbol{\theta}_s)|/\tau_s\} \right.$$
$$\left. \leq \frac{(21b - 10)^2}{144b(21b - 8)^2} \left[\sqrt{\frac{1}{9b^2} + \frac{n}{2u^2[\log(3J_t(J_t + 1)) + \log(nT)]}} - \frac{1}{3b} \right] \right\}. \tag{3.6}$$

Theorem 3. *Suppose there exists a constant $S \geq 1$ such that $p(a_t|h_t) \geq S^{-1}$ for all (h_t, a_t) pairs for $t = 1, \ldots, T$, and the margin condition (3.3) holds for some $C > 0$, $\alpha \geq 0$ and all positive ϵ_t for $t = 1, \ldots, T$. Assume assumptions (B1)–(B4) hold. Suppose the tuning parameters $\lambda_t, t = 1, \ldots, T$, satisfy*

$$\lambda_t \geq 96\sqrt{2}\big[1 + 2(T-t)\big] b \max\{c, \sigma, \eta\} \sqrt{\frac{[\log(12J_t) + \log(nT)]}{n}}, \tag{3.7}$$

and $\lambda_t^2 \geq c_{t,s}\lambda_s^2$ for $t = 1, \ldots, T$, $s = t, \ldots, T$, where $c_{t,t} = 1, c_{t,s} = 2(2\gamma + 5)(5S + 3)(T - t)^2 c_{t+1,s}/9$. Let Θ_t be the set defined in (3.6) and assume Θ_t is nonempty for $t = 1, \ldots, T$. Then for any $\sqrt{n} \geq ku \max_t\{\sqrt{\log(12J_t) + \log(nT)}\}$ where $k = 4u/(3\sqrt{2}b)$, with the probability at least $1 - 1/n$ we have

$$V(\pi^o) - V(\hat{\pi}) \leq \sum_{t=1}^{T} C_{1,t} \left[\min_{\boldsymbol{\theta}_t \in \Theta_t} \left(E\left[\Phi_t^\top \boldsymbol{\theta}_t - Q_t^o\right]^2 + K_{t1} \max_{s \in \{t, \ldots, T\}} \left\{ c_{t,s} \frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \right) \right]^{(1+\alpha)/(2+\alpha)},$$

where $C_{1,t} = [2^{2+3\alpha} S^{(1+\alpha)t} C]^{1/(2+\alpha)}$ and $K_{t1} = [64(105b - 38)^2]/[81(21b - 8)^2] + [32b(105b - 38)]/[3(21b - 8)(21b - 10)]$.

Furthermore, suppose $E[\Phi_{t2}^T(H_t, A_t)|H_t] = \mathbf{0}$ a.s. for $t = 1, \ldots, T$. Then with the probability at least $1 - 1/n$,

$$V(\pi^o) - V(\hat{\pi}) \leq \sum_{t=1}^{T} C_{1,t} \left[\min_{\boldsymbol{\theta}_t \in \Theta_t} \left(E\left[\Phi_{t2}^\top \boldsymbol{\theta}_{t2} - U_t^o\right]^2 + K_{t2} \max_{s \in \{t, \ldots, T\}} \left\{ \bar{c}_{t,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \right) \right]^{(1+\alpha)/(2+\alpha)},$$

where $C_{1,t} = [2^{2+3\alpha} S^{(1+\alpha)t} C]^{1/(2+\alpha)}$, $K_{t2} = [3 - (21b - 10)^2/[9(21b - 8)^2]]^2 + [2b/(21b - 8)][81(21b-8)^2/(21b-10)^2 - 3]$, $\bar{c}_{t,t} = 1$, and $\bar{c}_{t,s} = 2(T-t)^2(S+1)[81 \max_{s \in \{t+1, \ldots, T\}}\{\bar{c}_{t+1,s}/c_{t+1,s}\}/(16(1 - 2\gamma)^2) + 1][3 - (1 - 2\gamma)^2/9]\bar{c}_{t+1,s}$ for $t = 1, \ldots, T$, $s = t + 1, \ldots, T$.

The result follows from the inequalities (3.4) and (3.5) in Lemma 1 and inequalities (B.5) and (B.6) in Theorem 4 with $\varphi = \log(nT)$ and $\gamma = 1/(21b - 8)$.

Remark. Assumptions (B1)–(B4) are employed in the proof of the finite sample upper bound results. Assumption (B1) implies that the error terms do not have heavy tails. Assumptions (B1) and (B2) are needed to show that the sample mean is concentrated around the true mean. Assumption (B3) is used to avoid collinearity. In addition, for any $\boldsymbol{\theta}_t, \acute{\boldsymbol{\theta}}_t \in \mathbb{R}^{J_t}$, one can easily verify that $E[\Phi_t^\top(\acute{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 = (W_t(\acute{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t))^\top M_t W_t(\acute{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t) \geq \tau_t(\sum_{j \in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\acute{\theta}_{tj} - \theta_{tj}|)^2/|I_t(\boldsymbol{\theta}_t)|$ by eigendecomposition and simple algebra, where $W_t = \text{diag}\{\bar{w}_{t1}, \ldots, \bar{w}_{tJ_t}\}$ and M_t is the gram-matrix which is provided in Assumption (B3). Thus, Assumption (B3) is a sufficient condition for

$$E[\Phi_t^\top(\acute{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2|I_t(\boldsymbol{\theta}_t)| \geq \tau_t(\sum_{j \in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\acute{\theta}_{tj} - \theta_{tj}|)^2, \tag{3.8}$$

for any $\boldsymbol{\theta}_t, \acute{\boldsymbol{\theta}}_t \in \mathbb{R}^{J_t}$, see, e.g., van de Geer (2008) for more details. Condition (3.8) is employed in the proofs of Lemmas 4 and 5 in the supplementary materials. This condition holds if the correlation $|E\phi_{tj}\phi_{tk}|/(\bar{w}_{tj}\bar{w}_{tk})$ is small for all $k \in I_t(\boldsymbol{\theta}_t)$, $j \neq k$. Assumption (B4) is used to ensure

$$\max_j \left| |E[\Phi_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)\phi_{tj}/\bar{w}_{tj}] \right| \leq \gamma\lambda_t b, \tag{3.9}$$

for $\boldsymbol{\theta}_t \in \Theta_t$ to derive Theorems 4 in the Appendix. When $\bar{w}_{tj} = (E\phi_{tj}^2)^{1/2}$ as in Qian and Murphy (2011), condition (3.9) is satisfied with $b = 1$.

3.4 Simulation

To assess the performance of each method, we conduct a set of numerical studies with two decision points. A variety of scenarios are considered with correctly specified (Scenarios 1-2) and misspecified (Scenario 3) models. In Scenarios 1-2, the simulated data are generated from

the model

$$Y = A_1 A_2 + A_2(\boldsymbol{\beta}_2^\top O_1 + O_2) + A_1(\boldsymbol{\beta}_1^\top O_1) + \varepsilon,$$

where $\varepsilon \sim N(0, 1^2)$. Treatments, A_1 and A_2, are randomly generated from $\{0, 1\}$ with equal probability 0.5. Covariates O_1 are generated from $N(0, I_p)$, and the intermediate covariate O_2 is generated by $O_2 = O_{11} + A_1 + A_1 O_{11} + e$, where $e \sim (0, 0.5^2)$. Then, the optimal treatment regime at stage 2 is $I(A_1 + \boldsymbol{\beta}_2^\top O_1 + O_2 > 0)$. Following this optimal treatment regime at stage 2, the Q-function at stage 1 is

$$\begin{aligned} Q_1(O_1, A_1) &= E[(A_1 + \boldsymbol{\beta}_2^\top O_1 + O_2)_+ | \, O_1, A_1] + A_1(\boldsymbol{\beta}_1^\top O_1) \\ &= \frac{1}{2\sqrt{2\pi}} \exp(-2\mu^2) + \mu\Phi(2\mu) + A_1(\boldsymbol{\beta}_1^\top O_1), \end{aligned}$$

where $z_+ = (|z| + z)/2$, $\mu = 2A_1 + \boldsymbol{\beta}_1^\top O_1 + O_{11} + A_1 O_{11}$, and $\Phi(z)$ is the cumulative distribution function of a standard normal distribution. Thus, the optimal treatment regime at state 1 is $I(Q_1(O_1, 1) - Q_1(O_1, 0) > 0)$, and we consider the following two scenarios with different degrees of sparsity level:

Scenario 1: $\boldsymbol{\beta}_1 = \mathbf{0}_p$, $\boldsymbol{\beta}_2 = (0, 0, 1, -1, \mathbf{0}_{p-4})$;

Scenario 2: $\boldsymbol{\beta}_1 = (\mathbf{0}_4, 1, -1, \mathbf{0}_{p-6})$, $\boldsymbol{\beta}_2 = (0, 0, 1, -1, \mathbf{0}_{p-4})$.

In Scenario 1, the active variables are $(A_1, O_{13}, O_{14}, O_2)$ at stage 2 and (O_{11}, O_{13}, O_{14}) at stage 1, since these are associated with A_2 and A_1, respectively. Similarly, in Scenario 2, the active variables are $(A_1, O_{13}, O_{14}, O_2)$ at stage 2 and $(O_{11}, O_{13}, O_{14}, O_{15}, O_{16})$ at stage 1.

We also consider Scenario 3 where the models are likely to be misspecified. In Scenario 3, treatments and baseline covariates are generated as in Scenarios 1-2; however, the generating models are different in Scenario 3 as below:

Scenario 3: Stage 1 outcome is generated from the model $Y_1 = 0.5\,O_{13}A_1 + \varepsilon_1$, where $\varepsilon_1 \sim N(0, 1^2)$, and stage 2 outcome is generated from the model $Y_2 = [(O_{11}^2 + O_{12}^2 - 0.2)(0.5 - O_{11}^2 - O_{12}^2) + Y_1]A_2 + \varepsilon_2$, where $\varepsilon_2 \sim N(0, 1^2)$.

In each scenario, we vary the number of observations n from 50 to 150 with the number of the first-stage covariates $p = 60$. We implement our partial regularization via orthogonalization using adaptive Lasso method (denoted as PRO-aLasso) along with penalized A-learning (PAL, Shi et al., 2018a) and backward outcome weighted learning (BOWL, Zhao et al., 2015). For our proposed method, we consider a linear working model for the Q-function as in Section 3.2, and the following terms are not penalized in each stage: $(1, A_1)$ in the first stage and $(1, A_2, A_1A_2)$ in the second stage.

For the PAL method, we apply the same algorithm described in Shi et al. (2018a); e.g., a linear regression for the baseline mean function, a logistic regression for the propensity score model with the SCAD penalties, and the Dantzig selector (Candes & Tao, 2007) on the A-learning estimating equation for variable selection. For the implementation of BOWL, we follow the estimation algorithm introduced in Zhao et al. (2015) with the linear kernel and the hinge surrogate loss.

Table 3.1 summarizes the performances of the methods based on 1,000 replications. For each replication, we compute three statistics for each method: false positive (FP; the number of inactive variables incorrectly included in the model), false negative (FN; the number of active variables left out of the model), and value function of the estimated optimal treatment regime. The value function is assessed by the estimated optimal DTR using an independent test dataset with sample size of $10,000$. Both FP and FN of the PRO-aLasso method are always smaller than (or at least equal to) that of PAL. Furthermore, it can be seen that in all cases the PRO-aLasso outperforms its counterparts in terms of higher value function estimate and better selection performance. The value estimation improves as the sample size grows for both PRO-aLasso and PAL; however, it is worth noting that in Scenarios 1-2 the former gets very close to the true

Table 3.1: Simulation results based on 1,000 replications. The median number of inactive variables incorrectly selected in the model, denoted by FP, and the median number of active variables left out of the model, denoted by FN, are recorded along with the mean absolute deviation in parentheses. The mean of values is also reported with the standard deviation in parentheses.

			Stage 2		Stage 1	
n	Method	Value	FP	FN	FP	FN
Scenario 1						
50	Truth	2.27	0	0	0	0
	PRO-aLasso	2.09 (0.20)	2 (2.97)	1 (1.48)	1 (1.48)	2 (1.48)
	PAL	1.66 (0.39)	1 (1.48)	2 (1.48)	2 (1.48)	3 (0)
	BOWL	0.90 (0.23)	58 (0)	0 (0)	57 (0)	0 (0)
150	Truth	2.27	0	0	0	0
	PRO-aLasso	2.25 (0.02)	0 (0)	0 (0)	0 (0)	1 (1.48)
	PAL	2.20 (0.07)	0 (0)	1 (0)	1 (1.48)	2 (0)
	BOWL	0.96 (0.15)	58 (0)	0 (0)	57 (0)	0 (0)
Scenario 2						
50	Truth	2.45	0	0	0	0
	PRO-aLasso	2.07 (0.24)	2.5 (2.22)	1 (1.48)	2 (2.97)	3 (1.48)
	PAL	1.54 (0.41)	1 (1.48)	3 (1.48)	2 (1.48)	4 (1.48)
	BOWL	0.92 (0.24)	58 (0)	0 (0)	55 (0)	0 (0)
150	Truth	2.47	0	0	0	0
	PRO-aLasso	2.41 (0.02)	0 (0)	0 (0)	0 (0)	1 (1.48)
	PAL	2.31 (0.11)	0 (0)	1 (0)	1 (1.48)	2 (0)
	BOWL	0.99 (0.16)	58 (0)	0 (0)	55 (0)	0 (0)
Scenario 3						
50	Truth	7.32	0	0	0	0
	PRO-aLasso	6.40 (1.14)	0 (0)	3 (0)	0 (0)	1 (0)
	PAL	3.30 (1.80)	2 (1.48)	4 (0)	4 (1.48)	1 (0)
	BOWL	3.38 (1.15)	58 (0)	0 (0)	59 (0)	0 (0)
150	Truth	7.32	0	0	0	0
	PRO-aLasso	6.78 (0.00)	0 (0)	3 (0)	0 (0)	1 (0)
	PAL	4.99 (1.77)	1 (1.48)	4 (0)	6 (2.97)	1 (0)
	BOWL	2.93 (0.68)	58 (0)	0 (0)	59 (0)	0 (0)

optimal value. In Scenario 3 where the treatment effect is misspecified, the PRO-aLasso method still performs significantly better than others. Although PAL and BOWL also misspecify the relationship, the proposed method performs favorably against other methods since it prevents a model from both overshrinking and overfitting. BOWL method fails in all scenarios due to a very high FP.

3.5 Real Data Application

In this section, we apply our proposed method to a data from the coronary psychosocial evaluation studies (COPES) trial and the comparison of depression interventions after acute coronary syndrome (CODIACS) vanguard trial. The COPES trial was designed to increase the probability of demonstrating benefit from an intervention to reduce depressive symptoms in post-acute coronary syndrome (ACS) patients (Davidson et al., 2010). In the study, 157 post-ACS patients with elevated depressive symptoms were randomly assigned among intervention options. The CODIACS trial (Davidson et al., 2013) was conducted to determine the feasibility, efficacy, and costs of a centralized, stepped, patient preference-based depression care system for ACS patients. There were 724 Potential participants with ACS. Among them, 177 patients were found eligible, and 150 were enrolled and randomly allocated among interventions. In both trials, the participants were randomized to receive 6 months of treatment care, where the medication use is administered at baseline and 6 months during in-person interviews and at 2 and 4 months by telephone. Patients received a centralized problem-solving treatment (PST) therapist, medication, both, or neither as an intervention option.

We combine data from two trials and also combine 2 and 4 months to consider a two-stage problem; e.g., $T = 2$ stages (or decision points). The primary outcome of interest Y is change in Beck Depression Inventory (BDI) scores over 6 months. We drop the observations if either outcome is missing or treatments are missing. Thus, a total of 227 subjects are used in this analysis. The dummy variables, $A_t = (A_{t1}, A_{t2}, A_{t3})$, are created for the actions at each stage

$(t = 1, 2)$ as follows:

$$A_{t1} = \begin{cases} 1 & \text{if received both} \\ 0 & \text{otherwise} \end{cases} \quad A_{t2} = \begin{cases} 1 & \text{if received pst} \\ 0 & \text{otherwise} \end{cases} \quad A_{t3} = \begin{cases} 1 & \text{if received medication} \\ 0 & \text{otherwise} \end{cases}$$

We consider 9 baseline covariates including treatment preference, age, sex, hispanic, baseline actions, and baseline BDI score. Any baseline covariates with more than two possible levels are coded as a vector of dummy variables; thus, we have $O_1 \in \mathbb{R}^{13}$. We consider MACE (Major Adverse Cardiac Event) and an intermediate BDI score as the second-stage covariates $O_2 \in \mathbb{R}^2$. Here, MACE variable is recoded as 1 if the MACE occurred and the date of MACE is within 4 months from the enrollment date, and 0 otherwise. The intermediate reward at 4 months indicates if the intermediate BDI is reduced at least by 3 units:

$$Y_{4m} = \begin{cases} 1 & \text{if BDI score at baseline} - \text{BDI score at 4 months} > 3 \\ 0 & \text{if BDI score at baseline} - \text{BDI score at 4 months} \leq 3 \end{cases}$$

We apply various regularization and variable selection methods to estimate the optimal regime. The data is randomly split into a training set (75%), which is used to estimate the optimal regime, and a test set (25%), which is used to estimate the value of the estimated DTR. To implement Q-learning, we consider a linear working model for the Q-function of the form in Section 3.2. In particular, the PRO-aLasso does not place a penalty on the intercept and A_2 at stage 2, and similarly, no penalty is applied on the intercept and A_1 at stage 1. We compare our method with the adaptive Lasso and the ridge regression, both of which place penalties on all coefficients. We also consider the forward variable selection by the Akaike information criterion (AIC) for comparison.

Table 3.2 reports the estimated value and the size of DTR. The size of DTR is equivalent to the sum of two components: the number of non-zero coefficients involving A_2 at stage 2, and

the number of non-zero coefficients involving A_1 at stage 1. It is called the size of DTR since it specifies the number of input variables required to construct the optimal DTR. The last column of the table is the observed outcome where it captures the performance of randomly assigned interventions in the trial. In Table 3.2, it can be seen that the PRO-aLasso yield the highest estimated value with small DTR size. Both the adaptive Lasso and the ridge regression tend to over-shrink, whereas the forward selection tends to overfit. By taking advantage of the PRO-aLasso algorithm, it improves the treatment decisions and attains the right balance between overshrinkage and overfitting.

Table 3.2: Estimated value and size of DTR in parentheses using different methods.

PRO-aLasso	aLasso	Ridge	Forward	Observed
4.66 (6)	3.11 (0)	3.11 (216)	3.71 (18)	3.11

We illustrate the distribution of the estimated optimal DTR using each method. The PRO-aLasso always recommends medication as the second-stage optimal regime and PST as the first-stage optimal regime. On the contrary, the adaptive Lasso fails to provide a recommended regime, since all coefficients are shrunken towards zero in both stages. Likewise, the ridge regression fails to produce a recommendation because all the coefficients are very close zero. This strengthens the idea of not penalizing the main effects when the goal is to construct the optimal DTR. The forward selection recommends medication only in about 63%, following by neither (28%), both PST and medication (3%), and PST only (6%) in the second stage, and it always recommends PST in the first stage.

3.6 Discussion

Discovering the optimal DTRs using backward inductions is an appealing approach due to the ease of implementation. In this paper, we propose a penalized A-learning to construct the optimal DTR that would maximize the expected outcome if implemented. This methodology

places a Lasso-type penalty on some or all variables to find a model that is simple and has a good prediction accuracy. Another major advantage of the proposed approach is that it handles a problem with not only multiple decision points but also numerous treatment options. We also provide generalization error bounds of the estimated DTR through our proposed L_1-penalized A-learning. The error bounds do not depend on correct specification of the treatment by covariate interactions in the Q-functions. Specifically, we examine the relationship between value and the Q-functions, and then we provide a finite sample upper bound on the difference between the optimal value and the value obtained by the estimated DTR. This upper bound guarantees that the value of the estimated DTR will converge to the optimal value if the optimal Q-functions are consistently estimated. However, the theoretical foundation is based on continuous outcome. Thus, it remains an interesting task for future studies to generalize it to various types of data, including binary, ordinal, and censored outcome.

In practice, we have developed the multi-stage PRO-aLasso algorithm to impose an adaptive Lasso penalty only on a pre-specified partial set of variables for the construction of optimal DTR. A simulation study over different scenarios have shown that the PRO-aLasso produces higher values and better selection performances compared to other competing methods. In the real data analysis, the proposed method yields simpler regimes with higher values compared to its counterparts. It is crucial to recognize that the use of PRO technique mitigates the risk of overshrinkage which can occur in a completely data-driven regularization method. The optimal DTRs which are estimated from the stable and interpretable model will provide good guidance on medical practitioners and future studies.

Chapter 4. Conclusion and Future Work

This dissertation explores two research problems in Chapters 2 and 3. The first is to construct the individualized mobile health (mHealth) application recommender system, and the second is to estimate the optimal dynamic treatment regimes (DTRs) from a clinical trial study. Each of these projects described in the previous chapters was motivated by the needs and benefits of personalized adaptive interventions.

Mobile health interventions are increasingly used in research and health care. Despite its popularity and importance, the absence of suitable method of evaluating such interventions has been widely acknowledged. Our work was motivated by the problem of developing a recommender algorithm for mHealth apps in the IntelliCare ecosystem. In Chapter 2, we proposed a partial regularization via orthogonality using the adaptive Lasso (PRO-aLasso) to integrate domain knowledge into the model building procedure when developing optimal policy. Several theoretical results were derived to validate the proposed method, including variable selection consistency, asymptotic normality (even when the number of parameters diverges), and the rate of convergence of the value of the estimated policy. Furthermore, the proposed PRO technique is versatile and computationally efficient, and it can be implemented with other penalization approaches, although we opted to use the adaptive Lasso for its oracle property. We have demonstrated that the PRO-aLasso method correctly selects a set of important variables and is superior in terms of policy stability and higher value functions, compared to its counterparts. However, a realistic recommender system should provide sequential decision rules that adapt over time, and ideally account for cumulative information. This provided us with the possible avenues for extending this line of research to the multi-stage setting.

In many clinical trials, some patients respond differently to interventions, and what works for one group may not necessarily work for another. This illustrates the importance and benefit of individualized clinical decision making. In Chapter 3, we proposed a penalized A-learning procedure with a Lasso-type penalty for the construction of optimal DTR. We also established a finite sample upper bound on the difference in values between the optimal DTR and the estimated DTR. This upper bound guarantees that the value of the estimated DTR will converge to the optimal value if certain conditions are met. Through a series of simulations and a real calibration, we have shown that the proposed methodology is significant when it is desired to reduce the risk of over-shrinkage in a data-driven modeling method. Another advantage of our approach is that it handles a problem involving not only multiple stages (i.e., time points) but also numerous treatment options. To our knowledge, most of the studies have so far focused mainly on binary treatment regimes. In future research, we aim to handle other types of outcome data; count data that has an excess of zero counts or binary outcome is another interesting topic that needs further investigation. Additionally, it is worthwhile to develop the estimation of optimal DTRs in infinite horizon settings. Handling censoring or complex decision making (e.g., dosage level, timing of treatment) is another challenging issue. There are many interesting problems remained to be solved in the area of DTR, and we plan to explore some of them in greater depth in future.

References

Bellman, R. (1957). Dynamic programming. Princeton, Princeton University Press.

Blatt, D., Murphy, S. A., & Zhu, J. (2004). A-learning for approximate planning.

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, *35*(6), 2313–2351.

Carney, R. M., Freedland, K. E., & Steinmeyer, B. (2008). Depression and five year survival following acute myocardial infarction: A prospective study. *Journal of Affective Disorders*, *109*(1), 133–138.

Chakraborty, B., Murphy, S., & Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical methods in medical research*, *19*(3), 317–343.

Cheung, K., Ling, W., Karr, C., Weingardt, K., Schueller, S., & Mohr, D. (2018). Evaluation of a recommender app for apps for the treatment of depression and anxiety: An analysis of longitudinal user engagement. *Journal of the American Medical Informatics Association*, *25*(8), 955–962.

Christensen, H., Griffiths, K., & Farrer, L. (2009). Adherence in internet interventions for anxiety and depression. *Journal of Medical Internet Research*, *11*(2), e13.

Davidson, K. W., Bigger, J. T., Burg, M. M., Carney, R. M., Chaplin, W. F., Czajkowski, S., Dornelas, E., Duer-Hefele, J., Frasure-Smith, N., Freedland, K. E., Haas, D. C., Jaffe, A. S., Ladapo, J. A., Lesperance, F., Medina, V., Newman, J. D., Osorio, G. A., Parsons, F., Schwartz, J. E., ... Ye, S. (2013). Centralized, stepped, patient preference-based treatment for patients with post-acute coronary syndrome depression: Codiacs vanguard randomized controlled trial. *Journal of the American Medical Association Internal Medicine*, *173*(11), 997–1004.

Davidson, K. W., Rieckmann, N., Clemow, L., Schwartz, J. E., Shimbo, D., Medina, V., Albanese, G., Kronish, I., Hegel, M., & Burg, M. M. (2010). Enhanced depression care for patients with acute coronary syndrome and persistent depressive symptoms: Coronary psychosocial evaluation studies randomized controlled trial. *Archives of internal medicine*, *170*(7), 600–608.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*(4), 407–499.

Ellis, J. J., Eagle, K. A., Kline-Rogers, E. M., & Erickson, S. R. (2005). Depressive symptoms and treatment after acute coronary syndrome. *International journal of cardiology*, *99*(3), 443–447.

Ernst, D., Stan, G. B., Goncalves, J., & Wehenkel, L. (2006). Clinical data based optimal STI strategies for HIV: A reinforcement learning approach (p. 65-72). In Proceedings of the machine learning conference of Belgium; The Netherlands (Benelearn).

Ertefaie, A. (2014). Constructing dynamic treatment regimes in infinite horizon settings. *arXiv no. 1406.0764*.

Fan, A., Lu, W., & Song, R. (2016). Sequential advantage selection for optimal treatment regime. *Annals of Applied Statistics*, *10*(1), 32–53.

Fan, J., & Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, *57*(8), 5467–5484.

Fan, J., Xue, L., & zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, *42*(3), 819–849.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., & Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, *99*(467), 710–723.

Foster, J., Taylor, J., & Ruberg, S. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*(24), 2867–2880.

Friedman, J., Hastie, J., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *1*(2), 302–332.

Friedrich, M. J. (2017). Depression is the leading cause of disability around the world. *Journal of the American Medical Association*, *317*(15), 1517.

Gey, S. (2012). Risk bounds for cart classifiers under a margin condition. *Pattern Recognition*, *45*(9), 3523–3534.

Gunter, L., Zhu, J., & Murphy, S. (2011). Variable selection for qualitative interactions. *Statistical Methodology*, *8*(1), 42–55.

Hollis, C., Falconer, C. J., Martin, J. L., Whittington, C., Stockton, S., Glazebrook, C., & Davies, E. B. (2017). Annual research review: Digital health interventions for children and young people with mental health problems – a systematic and meta-review. *Journal of Child Psychology and Psychiatry*, *58*(4), 474–503.

Huang, J., & Xie, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. *In Asymptotics: Particles, Processes and Inverse Problems. IMS Lecture Notes, Monograph Series*, *55*, 149–166.

IQVIA. (2017). The growing value of digital health: Evidence and impact on human health and the healthcare system. *Institute Report*.

Jia, J., & Yu, B. (2008). On model selection consistency of elastic net when $p \gg n$. *Technical Report 756, Dept. Statistics, Univ. California, Berkeley*.

Jiang, R., Lu, W., Song, R., Hudgens, M., & Naprvavnik, S. (2017). Doubly robust estimation of optimal treatment regimes for survival data—with application to an HIV/AIDS study. *Annals of Applied Statistics*, *11*(3), 1763–1786.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–255.

Kaptein, K. I., de Jonge, P., van den Brink, R. H., & Korf, J. (2006). Course of depressive symptoms after myocardial infarction and cardiac prognosis: A latent class analysis. *Psychosom Med*, *68*(5), 662–668.

Kazdin, A., & Blase, S. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, *6*(1), 21–37.

Laber, E., Linn, K., & Stefanski, L. (2014). Interactive model building for Q-learning. *Biometrika*, *101*(4), 831–1321.

Laber, E., & Zhao, Y. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, *102*(3), 501–514.

Lattie, E., Schueller, S., Sargent, E., Stiles-Shields, C., Tomasino, K., Corden, M., Begale, M., Karr, C., & Mohr, D. (2016). Uptake and usage of IntelliCare: A publicly available suite of mental health and well-being apps. *Internet Interventions*, *4*(2), 152–158.

Lavori, P. W., Dawson, R., & Rush, A. J. (2000). Flexible treatment strategies in chronic disease: Clinical and research implications. *Biol Psychiatry*, *48*(6), 605–614.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*, 365–411.

Lipkovich, I., Dmitrienko, A., Denne, J., & Enas, G. (2011). Subgroup identification based on differential effect search: A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, *30*(21), 2601–2621.

Lipkovich, I., Dmitrienko, A., & Ralph, B. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, *36*(1), 136–196.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., & Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimes. *Statistics in Medicine*, *37*(26), 3776–3788.

Lu, W., Zhang, H., & Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, *22*(5), 493–504.

Luckett, D. J., Laber, E. B., nd D. M. Maahs, A. R. K., Mayer-Davis, E., & Kosorok, M. R. (2019). Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, 1–15.

Marin, F., Gonzalez-Conejero, R., Capranzano, P., Bass, T. A., Roldan, V., & Angiolillo, D. J. (2009). Pharmacogenetics in cardiovascular antithrombotic therapy. *Journal of the American College of Cardiology*, *54*(12), 1041–1057.

Mohr, D., Schueller, S., Montague, E., Burns, M., & Rashidi, P. (2014). The behavioral intervention technology model: An integrated conceptual and technological framework for eHealth and mHealth interventions. *Journal of Medical Internet Research*, *16*(6), e146.

Mohr, D., Schueller, S., Riley, W., Brown, C., Cuijpers, P., Duan, N., Kwansny, M., Stiles-Shields, C., & Cheung, K. (2015). Trials of intervention principles: Evaluation methods for evolving behavior intervention technologies. *Journal of Medical Internet Research*, *17*(7), e166.

Moodie, E. E. M., Richardson, T. S., & Stephens, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics*, *63*(2), 447–455.

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

Murphy, S. A. (2005). A generalization error for Q-learning. *Journal of Machine Learning Research*, *6*, 1073–1097.

Murphy, S. A., Oslin, D. W., Rush, A. J., & Zhu, J. (2007). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychoarmacology*, *32*(2), 257–262.

Ng, A., & Jordan, M. (2000). Pegasus: A policy search method for large mdps and pomdps (p. 406–415). *In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*.

Oh, E. J., Qian, M., Cheung, K., & Mohr, D. C. (2020). Building health application recommender system using partially penalized regression. *Statistical Modeling in Biomedical Research, Springer*, 105–123.

Pineau, J., Bellernare, M. G., Rush, A. J., Ghizaru, A., & Murphy, S. A. (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence*, *88*(Suppl 2), S52–S60.

Piper, W. E., Boroto, D. R., Joyce, A. S., McCallum, M., & Azim, H. F. A. (1995). Pattern of alliance and outcome in short-term individual psychotherapy. *Psychotherapy*, *32*(12), 639–647.

Qi, Z., & Liu, Y. (2018). D-learning to estimate optimal individual treatment rules. *Electronic Journal of Statistics*, *12*(2), 3601–3638.

Qian, M., & Murphy, S. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, *39*(2), 1180–1210.

Qian, M. (2010). Model selection and l_1 penalization for individualized treatment rules. *Doctoral dissertation, University of Michigan*.

research2gudiance. (2013). Mobile health market report 2013-2017: The commercialization of mHealth application (vol. 3). *Berlin: research2guidance*.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions (p. 189-326). New York, Springer. In Proceedings of the second seattle Symposium in Biostatistics.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560.

Rudin, C., & Ertekin, Ş. (2018). Learning customized and optimized lists of rules with mathematical programming. *Mathematical Programming Computation*, *10*(4), 659–702.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, *29*(4), 640–661.

Shen, J., Wang, L., & Taylor, J. M. G. (2017). Estimation of the optimal regime in treatment of prostate cancer recurrence from observational data using flexible weighting models. *Biometrics*, *73*, 635–645.

Shi, C., Fan, A., Song, R., & Lu, W. (2018a). High-dimensional A-learning for optimal dynamic treatment regimes. *Annals of Statistics*, *46*(3), 925–957.

Shi, C., Song, R., Lu, W., & fu, B. (2018b). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society: Series B*, *80*(4), 681–702.

Song, R., Korosok, M., Zeng, D., Zhao, Y., Laber, E., & Yuan, M. (2015a). On sparse representation for optimal individualized treatment selection with outcome weighted learning. *Stat*, *4*(1), 59–68.

Song, R., Wang, W., Zeng, D., & Kosorok, M. R. (2015b). Penalized Q-learning for dynamic treatment regimens. *Statistica Sinica*, *25*(3), 901–920.

Soubies, E., Blanc-Féraud, L., & Aubert, G. (2017). A unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization. *SIAM Journal on Optimization*, *27*(3), 2034–2060.

Su, X., Zhou, T., Yan, X., & Fan, J. (2008). Interaction trees with censored survival data. *International Journal of Biostatistics*, *4*(1), 1–26.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction (1st ed.). Cambridge, MIT Press.

Sutton, R. S., & Barto, A. G. (2017). Reinforcement learning: An introduction (2nd ed.). Cambridge, MIT Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, *58*, 267–288.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, *32*(1), 135–166.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, *36*(2), 614–645.

Wang, L., Zhou, Y., Song, R., & Sherwood, B. (2018). Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, *113*(523), 1243–1254.

Watkins, C. J. (1989). Learning from delayed rewards. England. Ph.D. thesis.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), 279–292.

WHO. (2017). Depression and other common mental disorders: Global health estimates. *Geneva: World Health Organization*, 1–24.

Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for Lasso penalized regression. *Annals of Applied Statistics*, *2*(1), 224–244.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, *68*, 49–67.

Yuan, M., & Lin, Y. (2007). On the nonnegative garrote estimator. *Journal of the Royal Statistical Society: Series B*, *69*(2), 143–161.

Zhang, B., Tsiatis, A., Davidian, M., Zhang, M., & Laber, E. (2012a). Estimating optimal treatment regimes from a classification perspective. *Stat*, *1*(1), 103–114.

Zhang, B., Tsiatis, A., Laber, E., & Davidian, M. (2012b). A robust method for estimating optimal treatment regimes. *Biometrics*, *68*(4), 1010–1018.

Zhang, B., Tsiatis, A., Laber, E., & Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, *100*(3), 681–694.

Zhang, B., & Zhang, M. (2017). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics*, *74*(3), 891–899.

Zhang, B., & Zhang, M. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics*, *74*(3), 891–899.

Zhang, C.-H. (2007). Penalized linear unbiased selection. *Department of Statistics and Bioinformatics, Rutgers University*, *3*.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*(2), 894–942.

Zhang, C.-H., & Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, *36*(4), 1567–1594.

Zhang, Y., Laber, E., Davidian, M., & Tsiatis, A. (2018). Estimation of optimal treatment regimes using lists. *Journal of the American Statistical Association*, *71*(4), 895–904.

Zhang, Y., Laber, E., Tsiatis, A., & Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, *71*(4), 895–904.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2567.

Zhao, Y., Zeng, D., Laber, E., & Korosok, M. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, *110*(510), 583–598.

Zhao, Y., Zeng, D., Rush, A., & Korosok, M. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, *107*(449), 1106–1118.

Zhao, Y., Zhu, R., Chen, G., & Zheng, Y. (2020). Constructing dynamic treatment regimes with shared parameters for censored data. *Statistics in Medicine*, *39*(9), 1250–1263.

Zhou, X., & Kosorok, M. R. (2017a). Augmented outcome-weighted learning for optimal treatment regimes. *arXiv:1711.10654*.

Zhou, X., & Kosorok, M. R. (2017b). Causal nearest neighbor rules for optimal treatment regimes. *arXiv:1711.08451*.

Zhou, X., Mayer-Hamblett, N., Khan, U., & Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, *112*(517), 169–187.

Zhu, R., Zeng, D., & Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, *110*(512), 1770–1784.

Zhu, W., Zeng, D., & Song, R. (2019). Proper inference for value function in high dimensional Q-learning for dynamic treatment regimes. *Journal of the American Statistical Association*, *114*(527), 1404–1417.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320.

Zou, H., & Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, *37*(4), 1733–1751.

Appendix A. Appendices to Chapter 2

The proofs of Theorems 1 and 2 rely on the following assumptions (A1)–(A6) and Lemmas 2 and 3.

Assumptions.

(A1) $\epsilon \overset{\Delta}{=} Y - \boldsymbol{Z}^T\boldsymbol{\alpha}_0 - \boldsymbol{X}^T\boldsymbol{\beta}_0$ has mean zero and finite variance σ^2, and is independent of $(\boldsymbol{Z}, \boldsymbol{X})$ with $E(|\epsilon|^{2+\kappa}) < \infty$ for some $\kappa > 0$.

(A2) $(\boldsymbol{Z}^T, \boldsymbol{X}^T)^T$ is uniformly bounded.

(A3) There exist positive constants b and B such that $b \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq B$, where $\lambda_{\min}(\boldsymbol{\Sigma})$ and $\lambda_{\max}(\boldsymbol{\Sigma})$ are the smallest and the largest eigenvalues of $\boldsymbol{\Sigma}$, respectively.

(A4) $\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F \to_p 0$, where $\|\cdot\|_F$ stands for the Frobenius norm.

(A5) $\frac{\log p}{\log n} \to \nu$ for some $0 \leq \nu < 1$.

(A6) $\lambda_n = o(\sqrt{n})$, $\quad \frac{\lambda_n}{\sqrt{n}} n^{((1-\nu)(1+\delta)-1)/2} \to \infty$.

Remark. Assumptions (A1) and (A2) are employed in the proof of the asymptotic results which are regular requirements. Assumption (A3) implies that the matrix $\boldsymbol{\Sigma}$ has a reasonably good behavior. Assumption (A4) is needed for the random design case. Also, note that the convergence in Frobenius norm implies the convergence in operator norm, which further guarantees the consistency of the sample eigenvalues. When $p^2/n \to 0$, this assumption is usually satisfied, see, e.g., the proof of Lemma 3.1 in Ledoit and Wolf (2004) for more details. Assumptions (A5) and (A6) are needed for the rate of convergence and the oracle properties. Following Zou and

Zhang (2009), we can fix $\delta = \lceil \frac{2\nu}{1-\nu} \rceil + 1$ to avoid the tuning on δ, and the oracle properties hold as long as $\delta > \frac{2\nu}{1-\nu}$.

Lemma 2. *Suppose assumptions (A1)–(A6) hold. Then under model (2.2), we have*

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \end{pmatrix} \right\|_2^2 = \mathrm{O}_P\left(\frac{p}{n}\right).$$

We derive the excess loss bound of the PRO-aLasso estimator in Lemma 2, which helps to prove the oracle properties. Under the regularity conditions, Lemma 2 also tells us that the PRO-aLasso estimator is a root-(n/p)-consistent estimator.

Lemma 3. *Let us write* $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_{0\mathcal{J}}, 0)$ *and define*

$$(\tilde{\boldsymbol{\alpha}}_n, \tilde{\boldsymbol{\beta}}_{n\mathcal{J}}) = \arg\min_{(\boldsymbol{\alpha},\boldsymbol{\beta})} nE_n\left(Y - (\boldsymbol{Z}, \boldsymbol{X}_{\mathcal{J}})^T(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T\right)^2 + \lambda_n \sum_{j\in\mathcal{J}} \hat{w}_j|\beta_j|.$$

Then with probability tending to 1, $(\tilde{\boldsymbol{\alpha}}_n, \tilde{\boldsymbol{\beta}}_{n\mathcal{J}}, 0)$ *is the solution to* (2.3).

Lemma 3 provides an asymptotic characterization for solving the PRO-aLasso criterion. Lemma 3 also shows that the PRO-aLasso estimator possesses the oracle properties in Theorem 1.

Proof of Theorem 1. Denote $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ for any $\boldsymbol{\alpha} \in \mathbb{R}^{p_2}$, $\boldsymbol{\beta} \in \mathbb{R}^{p_1}$, $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\alpha}}_n^T, \hat{\boldsymbol{\beta}}_n^T)^T$, and $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0^T, \boldsymbol{\beta}_0^T)^T$. Let $\boldsymbol{\Phi} = (\boldsymbol{Z}, \boldsymbol{X}) \in \mathbb{R}^{p\times p}$, where $p_1 + p_2 = p$. We first prove the model selection consistency part. Lemma 3 shows that the estimator minimizing the objective function (2.3) is equivalent to $(\tilde{\boldsymbol{\theta}}_{n\mathcal{S}}, 0)$. Thus, it suffices to show that $\Pr(\min_{s\in\mathcal{S}} |\tilde{\theta}_{ns}| > 0) \to 1$. Let $\eta = \min_{s\in\mathcal{S}} |\theta_{0s}|$. Note that

$$\min_{s\in\mathcal{S}} |\tilde{\theta}_{ns}| > \min_{s\in\mathcal{S}} |\theta_{0s}| - \left\|\tilde{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}\right\|_2.$$

By Lemma 2, it is straightforward that

$$\left\|\tilde{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}\right\|_2^2 = \mathrm{O}_P\left(\frac{p_2 + r}{n}\right).$$

Therefore it follows that

$$\min_{s\in\mathcal{S}} |\tilde{\theta}_{ns}| > \eta - \sqrt{\frac{p_2 + r}{n}}\mathrm{O}_P(1)$$

and finally $\Pr(\min_{s\in\mathcal{S}} |\tilde{\theta}_{ns}| > 0) \to 1$.

Now we show the asymptotic normality part. Note that from Lemma 3 the estimator $\tilde{\boldsymbol{\theta}}_{n\mathcal{S}}$ satisfies the following first order equation:

$$-2nE_n\left[\phi_s(Y - \boldsymbol{\Phi}_{\mathcal{S}}^T\tilde{\boldsymbol{\theta}}_{n\mathcal{S}})\right] + \lambda_n\hat{w}_s\mathrm{sgn}(\tilde{\theta}_{ns})I(s \in \mathcal{S} \setminus \{1,\ldots,p_2\}) = 0 \quad \text{for } s \in \mathcal{S}.$$

Since $\theta_{0s} = 0$ for $\forall s \in \mathcal{S}^c$ and $\epsilon = Y - \boldsymbol{\Phi}_{\mathcal{S}}^T\boldsymbol{\theta}_{0\mathcal{S}}$, this equation can be written

$$\begin{aligned}-2nE_n\left[\phi_s\boldsymbol{\Phi}_{\mathcal{S}}^T(\boldsymbol{\theta}_{0\mathcal{S}} - \tilde{\boldsymbol{\theta}}_{n\mathcal{S}})\right] - 2nE_n(\phi_s\epsilon)& \\ + \lambda_n\hat{w}_s\mathrm{sgn}(\tilde{\theta}_{ns})I(s \in \mathcal{S} \setminus \{1,\ldots,p_2\}) = 0 \quad \text{for } s \in \mathcal{S}.&\end{aligned}$$

Therefore, we have

$$\sqrt{n}\boldsymbol{\Sigma}_{\mathcal{S}}(\tilde{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}) = \sqrt{n}E_n(\boldsymbol{\Phi}_{\mathcal{S}}\epsilon) - \frac{\lambda_n}{2\sqrt{n}}\hat{\boldsymbol{w}}_{\mathcal{S}}\,\mathrm{sgn}(\tilde{\boldsymbol{\theta}}_{n\mathcal{S}})\boldsymbol{I}_{\mathcal{S}} + \sqrt{n}(\boldsymbol{\Sigma}_{\mathcal{S}} - \hat{\boldsymbol{\Sigma}}_{\mathcal{S}})(\tilde{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}})$$

where $\boldsymbol{I}_{\mathcal{S}} = (0_1\ldots,0_{p_2},1_{p_2+1},\ldots,1_{p_2+r})^T$.

Let $D_n = \sqrt{n}\psi^T\boldsymbol{\Sigma}_{\mathcal{S}}^{1/2}(\tilde{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}})$. Then $D_n = T_1 + T_2 + T_3$, where

$$\begin{aligned}T_1 &= \sqrt{n}\psi^T\boldsymbol{\Sigma}_{\mathcal{S}}^{-1/2}E_n(\boldsymbol{\Phi}_{\mathcal{S}}\epsilon),\\ T_2 &= -\frac{\lambda_n}{2\sqrt{n}}\psi^T\boldsymbol{\Sigma}_{\mathcal{S}}^{-1/2}\hat{\boldsymbol{w}}_{\mathcal{S}}\,\mathrm{sgn}(\tilde{\boldsymbol{\theta}}_{n\mathcal{S}})\boldsymbol{I}_{\mathcal{S}},\end{aligned}$$

$$T_3 = \sqrt{n}\psi^T \boldsymbol{\Sigma}_{\mathcal{S}}^{-1/2}(\boldsymbol{\Sigma}_{\mathcal{S}} - \hat{\boldsymbol{\Sigma}}_{\mathcal{S}})(\tilde{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}).$$

Using similar techniques as in the proof of Zou and Zhang (2009), we obtain $D_n \to_d N(0, \sigma^2)$. The result follows from Lemma 3 that with probability tending to 1, $\sqrt{n}\psi^T \boldsymbol{\Sigma}_{\mathcal{S}}^{1/2}(\hat{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}) = D_n \to_d N(0, \sigma^2)$.

Proof of Theorem 2. We show the convergence rate of the value function of the estimated policy. Observe that

$$\begin{aligned} E[\boldsymbol{\Phi}^T(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)]^2 &= E[\boldsymbol{\Phi}_{\mathcal{S}}^T(\hat{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}) + \boldsymbol{\Phi}_{\mathcal{S}^c}^T(\hat{\boldsymbol{\theta}}_{n\mathcal{S}^c} - \boldsymbol{\theta}_{0\mathcal{S}^c})]^2 \\ &\leq 2(\hat{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}})^T E\left[\boldsymbol{\Phi}_{\mathcal{S}}\boldsymbol{\Phi}_{\mathcal{S}}^T\right](\hat{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}) \\ &\qquad + 2E(\hat{\boldsymbol{\theta}}_{n\mathcal{S}^c} - \boldsymbol{\theta}_{0\mathcal{S}^c})^T E\left[\boldsymbol{\Phi}_{\mathcal{S}^c}\boldsymbol{\Phi}_{\mathcal{S}^c}^T\right](\hat{\boldsymbol{\theta}}_{n\mathcal{S}^c} - \boldsymbol{\theta}_{0\mathcal{S}^c}) \\ &\leq 2B\left\|\hat{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}\right\|_2^2 + 2B\left\|\hat{\boldsymbol{\theta}}_{n\mathcal{S}^c} - \boldsymbol{\theta}_{0\mathcal{S}^c}\right\|_2^2. \end{aligned}$$

By Theorem 1, it is true that for any a_n

$$P\left(a_n^{-1}\left\|\hat{\boldsymbol{\theta}}_{n\mathcal{S}^c} - \boldsymbol{\theta}_{0\mathcal{S}^c}\right\|_2^2 > \epsilon\right) \leq P(\exists\, s \in \mathcal{S}^c, \hat{\theta}_s \neq 0) \to 0,$$

and since $\left\|\hat{\boldsymbol{\theta}}_{n\mathcal{S}} - \boldsymbol{\theta}_{0\mathcal{S}}\right\|_2^2 = \mathrm{O}_P(\frac{p_2+r}{n})$, we have

$$\begin{aligned} E[\boldsymbol{\Phi}^T(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)]^2 &\leq \mathrm{O}_P\left(\frac{p_2+r}{n}\right) + \mathrm{o}_P(1) \\ &= \mathrm{O}_P\left(\frac{p_2+r}{n}\right). \end{aligned}$$

Hence, using Theorem 1 of Qian and Murphy (2011), we obtain

$$V(\pi_0) - V(\hat{\pi}) \leq E[\boldsymbol{\Phi}^T(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)]^2$$

$$\leq \mathrm{O}_P\left[\left(\frac{p_2+r}{n}\right)^{(1+\eta)/(2+\eta)}\right].$$

Proof of Lemma 2. The PRO-aLasso estimator minimizing the objective function (2.3) can be re-written as

$$\hat{\boldsymbol{\theta}}_n(\lambda_n) = \arg\min_{\boldsymbol{\theta}} nE_n(Y - \boldsymbol{\Phi}^T\boldsymbol{\theta})^2 + \lambda_n \sum_{s=p_2+1}^{p} \hat{w}_s|\theta_s|.$$

We know that for $s \in \{1\ldots, p\}$, the estimator $\hat{\boldsymbol{\theta}}_n(\lambda_n)$ satisfies the first order equation:

$$-2nE_n\left[\phi_s(Y - \boldsymbol{\Phi}^T\hat{\boldsymbol{\theta}}_n(\lambda_n))\right] + \lambda_n\hat{w}_s\mathrm{sgn}(\hat{\theta}_{ns}(\lambda_n))I(s \in \{p_2+1,\ldots,p\}) = 0,$$

where $\mathrm{sgn}(\theta_s) = 1$ if $\theta_s > 0$, $\mathrm{sgn}(\theta_s) = -1$ if $\theta_s < 0$, and $\mathrm{sgn}(\theta_s) \in [-1, 1]$ if $\theta_s = 0$ for any $\theta_s \in \mathbb{R}$. This implies

$$-2nE_n\left[\boldsymbol{\Phi}(Y - \boldsymbol{\Phi}^T\hat{\boldsymbol{\theta}}_n(\lambda_n))\right] + \lambda_n\hat{\boldsymbol{w}}\,\mathrm{sgn}(\hat{\boldsymbol{\theta}}_n(\lambda_n))\boldsymbol{I} = 0,$$

where $\boldsymbol{I} = (0_1,\ldots,0_{p_2},1_{p_2+1},\ldots,1_p)^T$. Note that assumption (A4) implies $\lambda_{\min}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \leq \lambda_{\max}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \to_p 0$. Then by the Courant-Fischer min-max Theorem, we have $\lambda_{\min}(\boldsymbol{\Sigma}) + \lambda_{\min}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \leq \lambda_{\min}(\hat{\boldsymbol{\Sigma}})$ and $\lambda_{\max}(\hat{\boldsymbol{\Sigma}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) + \lambda_{\max}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})$. This implies $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}) \to_p b$ and $\lambda_{\max}(\hat{\boldsymbol{\Sigma}}) \to_p B$, respectively, by assumption (A3). Note that the estimator $\hat{\boldsymbol{\theta}}_n(0)$ satisfies

$$-2nE_n\left[\boldsymbol{\Phi}(Y - \boldsymbol{\Phi}^T\hat{\boldsymbol{\theta}}_n(0))\right] = 0.$$

Therefore,

$$E_n\boldsymbol{\Phi}\boldsymbol{\Phi}^T(\hat{\boldsymbol{\theta}}_n(\lambda_n) - \hat{\boldsymbol{\theta}}_n(0)) = \frac{\lambda_n\hat{\boldsymbol{w}}\,\mathrm{sgn}(\hat{\boldsymbol{\theta}}_n(\lambda_n))\boldsymbol{I}}{2n},$$

which yields

$$\left\|\hat{\boldsymbol{\theta}}_n(0) - \hat{\boldsymbol{\theta}}_n(\lambda_n)\right\|_2^2 \leq \frac{\lambda_n^2(\sum_{s=p_2+1}^{p} \hat{w}_s^2)}{4n^2(\lambda_{\min}(\hat{\boldsymbol{\Sigma}}))^2},$$

since $(\mathrm{sgn}(\theta_s))^2 \leq 1$ for any $\theta_s \in \mathbb{R}$. Also, note that $\hat{\boldsymbol{\theta}}_n(\lambda_n) - \boldsymbol{\theta}_0 = (\hat{\boldsymbol{\theta}}_n(\lambda_n) - \hat{\boldsymbol{\theta}}_n(0)) + \hat{\boldsymbol{\Sigma}}^{-1} E_n \boldsymbol{\Phi}\boldsymbol{\epsilon}$. Then it follows that

$$\begin{aligned}
\left\|\hat{\boldsymbol{\theta}}_n(\lambda_n) - \boldsymbol{\theta}_0\right\|_2^2 &\leq 2\left\|\hat{\boldsymbol{\theta}}_n(\lambda_n) - \hat{\boldsymbol{\theta}}_n(0)\right\|_2^2 + 2\frac{\|E_n\boldsymbol{\Phi}\boldsymbol{\epsilon}\|_2^2}{(\lambda_{\min}(\hat{\boldsymbol{\Sigma}}))^2} \\
&\leq 2\frac{\lambda_n^2(\sum_{s=p_2+1}^{p} \hat{w}_s^2) + n^2\|E_n\boldsymbol{\Phi}\boldsymbol{\epsilon}\|_2^2}{n^2(\lambda_{\min}(\hat{\boldsymbol{\Sigma}}))^2}, \\
&\leq 2\frac{\lambda_n^2 p_1 + \|nE_n\boldsymbol{\Phi}\boldsymbol{\epsilon}\|_2^2}{n^2(\lambda_{\min}(\hat{\boldsymbol{\Sigma}}))^2},
\end{aligned}$$

where we set $\hat{w}_s = 1$ for all $s \in \{p_2 + 1, \ldots, p\}$ in the last inequality. Note that $E\|nE_n\boldsymbol{\Phi}\boldsymbol{\epsilon}\|_2^2 = E(\sum_{i=1}^{n} \boldsymbol{\Phi}_i\epsilon_i)^2 = n\sigma^2 E(\boldsymbol{\Phi}^T\boldsymbol{\Phi}) = n\sigma^2 \mathrm{Tr}(E(\boldsymbol{\Phi}^T\boldsymbol{\Phi})) = n\sigma^2 \mathrm{Tr}(\boldsymbol{\Sigma}) \leq n\sigma^2 p\lambda_{\max}(\boldsymbol{\Sigma})$. Thus, we have $\left\|\hat{\boldsymbol{\theta}}_n(\lambda_n) - \boldsymbol{\theta}_0\right\|_2^2 = \mathrm{O}_P(p/n)$.

Proof of Lemma 3. Denote $\tilde{\boldsymbol{\theta}}_{n\mathcal{S}} = (\tilde{\boldsymbol{\alpha}}_n^T, \tilde{\boldsymbol{\beta}}_{n\mathcal{J}}^T)^T$. We show that $(\tilde{\boldsymbol{\theta}}_{n\mathcal{S}}, 0)$ satisfies the KKT condition of (2.3) with probability tending to 1. It suffices to show that

$$\Pr(\exists\, s \in \mathcal{S}^c \mid -2nE_n\phi_s(Y - \boldsymbol{\Phi}_{\mathcal{S}}^T\tilde{\boldsymbol{\theta}}_{n\mathcal{S}})| > \lambda_n\hat{w}_s) \to 0. \tag{A.1}$$

Note that (A.1) can be regarded as the adaptive elastic-net problem when the L_2 penalty is eliminated, and thus the proof follows Zou and Zhang (2009).

Appendix B. Appendices to Chapter 3

B.1. Multi-stage PRO-aLasso Algorithm

We first divide Φ_t into two parts: those need to be penalized, denoted by $X_t \in \mathbb{R}^{p_{t1}}$, and those left unpenalized, denoted by Z_t. Usually $Z_t \in \mathbb{R}^{p_{t2}}$ is low-dimensional and only includes several key variables. In this case, we can consider a working model as

$$Q_t(H_t, A_t; \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = Z_t^\top(H_t, A_t)\boldsymbol{\alpha}_t + X_t^\top(H_t, A_t)\boldsymbol{\beta}_t, \tag{B.1}$$

where model (B.1) is equivalent to model (3.1) by letting $\Phi_t = (Z_t, X_t)$ and $\boldsymbol{\theta}_t = (\boldsymbol{\alpha}_t^\top, \boldsymbol{\beta}_t^\top)^\top$. The PRO method aims to minimize the following stage-t objective function

$$L_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \mathbb{E}_n[\tilde{Y}_t - Q_t(H_t, A_t; \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)]^2 + \lambda_t \sum_{j=1}^{p_{t1}} w_{tj}|\beta_{tj}|,$$

where $\tilde{Y}_T = Y_T$, $\tilde{Y}_t = Y_t + \sum_{s=t+1}^{T}\left[Y_s + \max_{a_s} Q_s(H_s, a_s; \hat{\boldsymbol{\alpha}}_s, \hat{\boldsymbol{\beta}}_s) - Q_s(H_s, A_s; \hat{\boldsymbol{\alpha}}_s, \hat{\boldsymbol{\beta}}_s)\right]$ for $t = T-1, \ldots, 1$, and λ_t is a tuning parameter controlling the amount of penalization at time t. Note that $\boldsymbol{w}_t = (w_{t1}, \ldots, w_{tp_{t1}})$ is a vector of weights adjusting a level of penalization on individual variables at time t. One can adopt $\hat{\boldsymbol{w}}_t = 1/|\tilde{\boldsymbol{\beta}}_t|^\delta$ for some $\delta > 0$ with $\tilde{\boldsymbol{\beta}}_t$ being a root-(n/p_{t1}) consistent estimator, which is known as the adaptive Lasso (aLasso). In practice, we propose to set $\tilde{\boldsymbol{\beta}}_t$ as perturbed elastic net estimates, following Zou and Zhang (2009), and 5-fold cross-validation can be used to select an optimal pair of (δ, λ_t). The multi-stage PRO-aLasso algorithm, which imposes an adaptive Lasso penalty only on X_t but not on Z_t, is described in Algorithm 2.

Algorithm 2 Multi-stage PRO-aLasso Algorithm

Input: data $(O_1, A_1, Y_1, \ldots, O_T, A_T, Y_T)$

Output: DTR $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_T)$

Set history by $H_1 = O_1$

for $t = 2, \ldots, T$ **do**

  Set history $H_t = (O_1, A_1, O_2, A_2, \ldots, O_t)$

end for

Set $\tilde{Y}_T = Y_T$

for $t = T, \ldots, 1$ **do**

  if $t \in \{T-1, \ldots, 1\}$ **then**

    Define $\tilde{Y}_t = Y_t + \sum_{s=t+1}^{T} \left[Y_s + \left(\max_{a_s} (X_s^\top \hat{\boldsymbol{\beta}}_s + Z_s^\top \hat{\boldsymbol{\alpha}}_s) \right) - (X_s^\top \hat{\boldsymbol{\beta}}_s + Z_s^\top \hat{\boldsymbol{\alpha}}_s) \right]$

  end if

  Formulate X_t and Z_t as a function of H_t and A_t to impose a penalty only on X_t but not on Z_t

  $\hat{\boldsymbol{\upsilon}}_t \leftarrow \arg\min_{\boldsymbol{\upsilon}_t} E_n \left(\tilde{Y}_t - Z_t^\top \boldsymbol{\upsilon}_t \right)^2$

  for $j = 1, \ldots, p_{t1}$ **do**

    $\hat{\boldsymbol{\gamma}}_{tj} \leftarrow \arg\min_{\boldsymbol{\gamma}_{tj}} E_n \left(X_{tj} - Z_t^\top \boldsymbol{\gamma}_{tj} \right)^2$

  end for

  $\hat{\boldsymbol{\Gamma}}_t \leftarrow (\hat{\boldsymbol{\gamma}}_{t1}, \ldots, \hat{\boldsymbol{\gamma}}_{tp_{t1}})$

  Construct $\hat{\boldsymbol{w}}_t = |\bar{\boldsymbol{\beta}}_t|^{-\delta}$ for some $\delta > 0$ with $\bar{\boldsymbol{\beta}}_t$ being a root-(n/p_{t1})-consistent estimator, which is

    obtained from the response $\tilde{Y}_t - Z_t^\top \hat{\boldsymbol{\upsilon}}_t$ and the predictor matrix $X_t - Z_t^\top \hat{\boldsymbol{\Gamma}}_t$

  Define $(X_t - Z_t^\top \hat{\boldsymbol{\Gamma}}_t)^* = (X_t - Z_t^\top \hat{\boldsymbol{\Gamma}}_t)/\hat{\boldsymbol{w}}_t$

  Solve the lasso problem for all λ_t,

$$\hat{\boldsymbol{\beta}}_t^* \leftarrow \arg\min_{\boldsymbol{\beta}} E_n \left(\tilde{Y}_t - Z_t^\top \hat{\boldsymbol{\upsilon}}_t - ((X_t - Z_t^\top \hat{\boldsymbol{\Gamma}}_t)^*)^\top \boldsymbol{\beta}_t \right)^2 + \lambda_t \sum_{j=1}^{p_{t1}} |\beta_{tj}|$$

  $\hat{\boldsymbol{\beta}}_t \leftarrow \hat{\boldsymbol{\beta}}_t^* / \hat{\boldsymbol{w}}_t$

  $\hat{\boldsymbol{\alpha}}_t \leftarrow \hat{\boldsymbol{\upsilon}}_t - \hat{\boldsymbol{\Gamma}}_t \hat{\boldsymbol{\beta}}_t$

  $\hat{\pi}_t \in \arg\max_{\boldsymbol{a}_t} (X_t^\top \hat{\boldsymbol{\beta}}_t + Z_t^\top \hat{\boldsymbol{\alpha}}_t)$

end for

$\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_T)$

B.2. Proof of Lemma 1

The proof of Lemma 1 follows that of Theorem II.1 in Qian (2010). □

B.3. Upper bounds for $E\left[\Phi_t^\top \hat{\theta}_t - Q_t^o\right]^2$ and $E\left[\Phi_{t2}^\top \hat{\theta}_{t2} - U_t^o\right]^2$

For any given $\varphi > 0$ and for $t = 1, \ldots, T$ define

$$\Theta_t = \left\{ \theta_t \in \Theta_t^* : \max_{s \in \{t,\ldots,T\}} \{|I_s(\theta_s)|/\tau_s\} \le \frac{(1-2\gamma)^2}{144b} \left[\sqrt{\frac{1}{9b^2} + \frac{n}{2u^2[\log(3J_t(J_t+1)) + \varphi]}} - \frac{1}{3b} \right] \right\}. \tag{B.2}$$

Theorem 4. *Suppose there exists a constant $S \ge 1$ such that $p(a_t|h_t) \ge S^{-1}$ for all (h_t, a_t) pairs for $t = 1, \ldots, T$. Assume assumptions (B1)–(B4) hold. For any given $0 \le \gamma < 2/(21b - 8)$ and $\varphi > 0$, suppose the tuning parameters $\lambda_t, t = 1, \ldots, T$, satisfy*

$$\lambda_T \ge \frac{8\max\{3c, 4\eta\}u[\log(12J_T) + \varphi]}{[1 - 2\gamma(3b-2)]n} + \frac{12\max\{\sigma, 2\eta\}b}{[1-2\gamma(3b-2)]}\sqrt{\frac{2[\log(12J_T) + \varphi]}{n}}, \tag{B.3}$$

$$\lambda_t \ge \max\left\{2c, \frac{8}{3}\left[1 + 2(T-t)\right]\eta\right\}\frac{u[\log(12J_t) + \varphi]}{\delta n} + \max\left\{\sigma, 2\left[1 + 2(T-t)\right]\eta\right\}\frac{b}{\delta}\sqrt{\frac{2[\log(12J_t) + \varphi]}{n}}, \tag{B.4}$$

and $\lambda_t^2 \ge c_{t,s}\lambda_s^2$ with $c_{t,t} = 1$, $c_{t,s} = \frac{2}{9}(2\gamma + 5)(5S + 3)(T - t)^2 c_{t+1,s}$,

for $t = 1, \ldots, T$, $s = t, \ldots, T$, where $\delta = (1 + 4\gamma)/12 - 7b\gamma/8$. Let Θ_t be the set defined in (B.2) and assume Θ_t is nonempty for $t = 1, \ldots, T$. Then with the probability at least $1 - T\exp(-\varphi)$,

we have

$$E[\Phi_t^\top \hat{\boldsymbol{\theta}}_t - Q_t^o]^2 \le \min_{\boldsymbol{\theta}_t \in \Theta_t} \left(E[\Phi_t^\top \boldsymbol{\theta}_t - Q_t^o]^2 + K_{t1} \max_{s \in \{t,\ldots,T\}} \left\{ c_{t,s} \frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \right), \tag{B.5}$$

where $K_{t1} = [64(2\gamma+5)^2]/81 + [32\gamma b(2\gamma+5)]/[3(1-2\gamma)]$.

Furthermore, suppose $E[\Phi_{t2}^T(H_t, A_t)|H_t] = \mathbf{0}$ *a.s. Then with probability at least* $1 - T\exp(-\varphi)$,

$$E[\Phi_{t2}^\top \hat{\boldsymbol{\theta}}_{t2} - U_t^o]^2 \le \min_{\boldsymbol{\theta}_t \in \Theta_t} \left(E[\Phi_{t2}^\top \boldsymbol{\theta}_{t2} - U_t^o]^2 + K_{t2} \max_{s \in \{t,\ldots,T\}} \left\{ \bar{c}_{t,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \right), \tag{B.6}$$

where $K_{t2} = [3 - [(1-2\gamma)^2]/9]^2 + 2\gamma b[81/[(1-2\gamma)^2] - 3]$, $\bar{c}_{t,t} = 1$, *and*

$$\bar{c}_{t,s} = 2(T-t)^2(S+1)\left[\frac{81 \max_{s \in \{t+1,\ldots,T\}}\{\bar{c}_{t+1,s}/c_{t+1,s}\}}{16(1-2\gamma)^2} + 1\right]\left[3 - \frac{(1-2\gamma)^2}{9}\right]\bar{c}_{t+1,s},$$

for $t = 1, \ldots, T$, $s = t+1, \ldots, T$.

Proof of Theorem 4. For any $\boldsymbol{\theta}_t \in \Theta_t, t = 1, \ldots, T$, we denote

$$\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) = Y_t + \sum_{s=t+1}^{T} \left[Y_s + \max_{a_s} \Phi_s^T(H_s, a_s)\boldsymbol{\theta}_s - \Phi_s^T(H_s, A_s)\boldsymbol{\theta}_s \right], \tag{B.7}$$

when $t = T-1, \ldots, 1$, and $\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) \equiv Y_T$ when $t = T$ for the convenience of notation. Denote $|\mathscr{A}_t|$ the number of treatment options at stage t. Define the events

$$\Omega_{t,1}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T) = \left\{ \max_{j,k \in \{1,\ldots,J_t\}} \left| (E - \mathbb{E}_n)\left(\frac{\phi_{tj}\phi_{tk}}{\bar{w}_{tj}\bar{w}_{tk}} \right) \right| \le \frac{(1-2\gamma)^2}{144 \max_{s \in \{t,\ldots,T\}}\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}} \right\},$$

$$\Omega_{t,2}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T) = \left\{ \max_{j \in \{1,\ldots,J_t\}} \left| \mathbb{E}_n\left[\left(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) - \Phi_t^T\boldsymbol{\theta}_t \right) \frac{\phi_{tj}}{\bar{w}_{tj}} \right] \right| \le \frac{4\gamma+1}{6}\lambda_t \right\},$$

$$\Omega_{t,3}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T) = \left\{ \max_{j,k \in \{1,\ldots,J_t\}} \left| (E - \mathbb{E}_n)\left(\sum_{a_t \in \mathscr{A}_t} \frac{\phi_{tj}(H_t, a_t)\phi_{tk}(H_t, a_t)}{\bar{w}_{tj}\bar{w}_{tk}} \right) \right| \right.$$

$$\leq \frac{(1-2\gamma)^2|\mathscr{A}_t|}{144\max_{s\in\{t,\ldots,T\}}\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}}\Bigg\}.$$

Note that by Cauchy-Schwarz inequality,

$$\max_j \left|E\left[\Phi_t^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}_t^*)\frac{\phi_{tj}}{\bar{w}_{tj}}\right]\right| \leq \sqrt{E[\Phi_t^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}_t^*)]^2 \max_j E[\phi_{tj}/\bar{w}_{tj}]^2} \leq \gamma\lambda_t b,$$

where the second inequality holds from Assumption (B4). Thus for $\boldsymbol{\theta}_t \in \Theta_t$,

$$\begin{aligned}
&E[\Phi_t^\top\hat{\boldsymbol{\theta}}_t - Q_t^o]^2 \\
&\quad= E[\Phi_t^\top\boldsymbol{\theta}_t - Q_t^o]^2 + E[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]^2 + 2E[\Phi_t^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}_t^*)][\Phi_t^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)] \\
&\quad\leq E[\Phi_t^\top\boldsymbol{\theta}_t - Q_t^o]^2 + E[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]^2 + 2\max_{j\in\{1,\ldots,J_t\}}\left|E\left[\Phi_t^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}_t^*)\frac{\phi_{tj}}{\bar{w}_{tj}}\right]\right|\Big(\sum_{j=1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big) \\
&\quad\leq E[\Phi_t^\top\boldsymbol{\theta}_t - Q_t^o]^2 + E[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]^2 + 2\gamma\lambda_t b\Big(\sum_{j=1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big).
\end{aligned}$$

By Lemma 4 below, we have on the event $\Omega_{t,1}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)$ for $t=T$, and on the event $\Omega_{t,1}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\cap\left(\cap_{s=t+1}^T\Omega_{s,3}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T)\right)$ for $t<T$ that

$$E[\Phi_t^\top\hat{\boldsymbol{\theta}}_t - Q_t^o]^2 \leq \min_{\boldsymbol{\theta}_t\in\Theta_t}\left(E[\Phi_t^\top\boldsymbol{\theta}_t - Q_t^o]^2 + K_{t1}\max_{s\in\{t,\ldots,T\}}\left\{c_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}\right),$$

where $K_{t1} = [64(2\gamma+5)^2]/81 + [32\gamma b(2\gamma+5)]/[3(1-2\gamma)]$.

If $E[\Phi_{t2}^T(H_t,A_t)|H_t] = \mathbf{0}$ a.s., by Lemma 5 we have on the event $\Omega_{t,1}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)$ for $t=T$, and on the event $\Omega_{t,1}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\cap\left(\cap_{s=t+1}^T\Omega_{s,3}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T)\right)$ for $t<T$ that

$$\begin{aligned}
&E[\Phi_{t2}^\top\hat{\boldsymbol{\theta}}_{t2} - U_t^o]^2 \\
&\quad\leq E[\Phi_{t2}^\top\boldsymbol{\theta}_{t2} - U_t^o]^2 + E[\Phi_{t2}^\top(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})]^2 + 2\gamma\lambda_t b\Big(\sum_{j=J_{t1}+1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)
\end{aligned}$$

$$\leq \min_{\boldsymbol{\theta}_t \in \Theta_t} \left(E[\Phi_{t2}^\top \boldsymbol{\theta}_{t2} - U_t^o]^2 + K_{t2} \max_{s \in \{t,\ldots,T\}} \left\{ \bar{c}_{t,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \right),$$

where $K_{t2} = [3 - [(1-2\gamma)^2]/9]^2 + 2\gamma b[81/[(1-2\gamma)^2] - 3]$.

The conclusion of the theorem follows from the union probability bounds of the events $\Omega_{t,1}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)$, $\Omega_{t,2}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)$, and $\Omega_{s,3}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T)$ for $s = t+1, \ldots, T$ provided in Lemmas 6, 7, and 8. □

Lemma 4. *Suppose there exists a constant $S \geq 1$ such that $p(a_t|h_t) \geq S^{-1}$ for all (h_t, a_t) pairs, and Assumption (B3) holds. Assume $\lambda_t^2 \geq c_{t,s}\lambda_s^2$ for $t = 1, \ldots, T$, $s = t, \ldots, T$, where $c_{t,t} = 1, c_{t,s} = 2(2\gamma+5)(5S+3)(T-t)^2 c_{t+1,s}/9$. Then, on the event $\cap_{t=1}^T \left\{ \Omega_{t,1}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{t,2}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{t+1,3}(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) \right\}$, we have*

$$\sum_{j=1}^{J_t} \bar{w}_{tj} |\hat{\theta}_{tj} - \theta_{tj}| \leq \frac{16(2\gamma+5)}{3(1-2\gamma)\lambda_t} \max_{s \in \{t,\ldots,T\}} \left\{ c_{t,s} \frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \tag{B.8}$$

$$E[\Phi_t^T(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 \leq \frac{64(2\gamma+5)^2}{81} \max_{s \in \{t,\ldots,T\}} \left\{ c_{t,s} \frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\}, \tag{B.9}$$

for any $\boldsymbol{\theta}_t \in \Theta_t$, $t = 1, \ldots, T$, and $\Omega_{T+1,3}(\boldsymbol{\theta}_{T+1})$ is defined as the universe for the convenience of notation.

Lemma 5. *Suppose all conditions in Lemma 4 hold. Assume $E[\Phi_{t2}^\top(H_t, A_t)|H_t] = \mathbf{0}$ a.s. for $t = 1, \ldots, T$. Then on the event $\cap_{t=1}^T \left\{ \Omega_{t,1}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{t,2}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{t+1,3}(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) \right\}$, we have*

$$\sum_{j=J_{t1}+1}^{J_t} \bar{w}_{tj} |\hat{\theta}_{tj} - \theta_{tj}| \leq \left[\frac{81}{(1-2\gamma)^2} - 3 \right] \lambda_t^{-1} \max_{s \in \{t,\ldots,T\}} \left\{ \bar{c}_{t,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \tag{B.10}$$

$$\text{and } E[\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})]^2 \leq \left[3 - \frac{(1-2\gamma)^2}{9} \right]^2 \max_{s \in \{t,\ldots,T\}} \left\{ \bar{c}_{t,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\}, \tag{B.11}$$

for any $\boldsymbol{\theta}_t \in \Theta_t$, $t = 1, \ldots, T$, *where* $c_{t,s}$ *is defined in Lemma 4,* $\bar{c}_{t,t} = 1$ *and*

$$\bar{c}_{t,s} = 2(T-t)^2(S+1)\left[\frac{81 \max_{s \in \{t+1,\ldots,T\}}\{\bar{c}_{t+1,s}/c_{t+1,s}\}}{16(1-2\gamma)^2} + 1\right]\left[3 - \frac{(1-2\gamma)^2}{9}\right]\bar{c}_{t+1,s},$$

for $t = 1, \ldots, T$, $s = t+1, \ldots, T$.

Lemma 6. *Suppose Assumptions (B2) and (B4) hold. Then for any* $\varphi > 0$ *and* $\boldsymbol{\theta}_t \in \Theta_t$, $\mathbf{P}(\{\Omega_{t,1}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)\}^C) \leq \exp(-\varphi)/3$ *for* $t = 1, \ldots, T$.

Lemma 7. *Suppose Assumptions (B1), (B2), and (B4) hold. Then for any* $\varphi > 0$, *if* λ_t *satisfies conditions* (B.3), (B.4) *and* $\lambda_t^2 \geq c_{t,s}\lambda_s^2$, *where* $c_{t,s}$ *is defined in Lemma 4, then for* $\boldsymbol{\theta}_t \in \Theta_t$, $\mathbf{P}(\{\Omega_{t,2}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)\}^C) \leq \exp(-\varphi)/3$ *for* $t = 1, \ldots, T$.

Lemma 8. *Suppose Assumptions (B2) and (B4) hold. Then for any* $\varphi > 0$ *and* $\boldsymbol{\theta}_t \in \Theta_t$, $\mathbf{P}(\{\Omega_{t,3}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)\}^C) \leq \exp(-\varphi)/3$ *for* $t = 1, \ldots, T$.

Supplementary Materials

Proof of Lemma 4.

We use induction to prove the results. At the last stage T, note that the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_T$ satisfies the following first order condition:

$$-2\mathbb{E}_n[(Y_T - \Phi_T^T\hat{\boldsymbol{\theta}}_T)\phi_{Tj}] + \lambda_T w_{Tj}\mathrm{sgn}(\hat{\theta}_{Tj}) = 0 \text{ for } j = 1, \ldots, J_T,$$

where $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = -1$ if $x < 0$ and $\mathrm{sgn}(x) \in [-1, 1]$ if $x = 0$ for any $x \in \mathbb{R}$. This implies

$$-2\mathbb{E}_n[(Y_T - \Phi_T^T\hat{\boldsymbol{\theta}}_T)\Phi_T^T\boldsymbol{\theta}_T] + \lambda_T \sum_{j=1}^{J_T} w_{Tj}\mathrm{sgn}(\hat{\theta}_{Tj})\theta_{Tj} = 0$$

for any $\boldsymbol{\theta}_T \in \mathbb{R}^{J_T}$. In particular, $-2\mathbb{E}_n[(Y_T - \Phi_T^T\hat{\boldsymbol{\theta}}_T)\Phi_T^T\hat{\boldsymbol{\theta}}_T] + \lambda_T \sum_{j=1}^{J_T} w_{Tj}|\hat{\theta}_{Tj}| = 0$. Therefore, for any $\boldsymbol{\theta}_T \in \mathbb{R}^{J_T}$, we have

$$\begin{aligned} 0 &= 2\mathbb{E}_n[(Y_T - \Phi_T^T\hat{\boldsymbol{\theta}}_T)\Phi_T^T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T)] + \lambda_T \sum_{j=1}^{J_T} w_{Tj}\mathrm{sgn}(\hat{\theta}_{Tj})\theta_{Tj} - \lambda_T \sum_{j=1}^{J_T} w_{Tj}|\hat{\theta}_{Tj}| \\ &\leq 2\mathbb{E}_n[(Y_T - \Phi_T^T\hat{\boldsymbol{\theta}}_T)\Phi_T^T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T)] + \lambda_T \sum_{j=1}^{J_T} w_{Tj}|\theta_{Tj}| - \lambda_T \sum_{j=1}^{J_T} w_{Tj}|\hat{\theta}_{Tj}|. \end{aligned} \tag{S.1}$$

Fix n. Following (S.1), on the event $\Omega_{T,2}(\boldsymbol{\theta}_T)$, we have

$$\begin{aligned} 0 \leq\ & 2 \max_{j\in\{1,\ldots,J_T\}} \left|\mathbb{E}_n\left[(Y_T - \Phi_T^T\boldsymbol{\theta}_T)\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| \left(\sum_{j=1}^{J_T} \bar{w}_{Tj}|\hat{\theta}_{Tj} - \theta_{Tj}|\right) - 2\mathbb{E}_n[\Phi_T^T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T)]^2 \\ &+ \lambda_T \sum_{j\in I_T(\boldsymbol{\theta}_T)} \bar{w}_{Tj}|\hat{\theta}_{Tj} - \theta_{Tj}| - \lambda_T \sum_{j\in I_T^c(\boldsymbol{\theta}_T)} \bar{w}_{Tj}|\hat{\theta}_{Tj}| \end{aligned}$$

$$\leq \frac{4(\gamma+1)}{3}\lambda_T\Big(\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)-\frac{2(1-2\gamma)}{3}\lambda_T\Big(\sum_{j\in I_T^c(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}|\Big)$$
$$-2\mathbb{E}_n[\Phi_T^T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T)]^2. \quad \text{(S.2)}$$

This implies

$$\sum_{j\in I_T^c(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}| \leq \frac{2(\gamma+1)}{1-2\gamma}\Big(\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big) \quad \text{(S.3)}$$
$$\text{and } \mathbb{E}_n[\Phi_T^T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T)]^2 \leq \frac{2(\gamma+1)}{3}\lambda_T\Big(\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big). \quad \text{(S.4)}$$

Using (S.3) and condition (3.8) on the event $\Omega_{T,1}(\boldsymbol{\theta}_T)$, we have

$$\begin{aligned}
&-\mathbb{E}_n[\Phi_T^T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T)]^2 \\
&\quad\leq \max_{j,k\in\{1,\ldots,J_T\}}\Big|(E-\mathbb{E}_n)\Big(\frac{\phi_{Tj}\phi_{Tk}}{\bar{w}_{Tj}\bar{w}_{Tk}}\Big)\Big|\Big(\sum_{j=1}^{J_T}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)^2 - E[\Phi_T^T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T)]^2 \\
&\quad\leq \frac{\tau_T}{16|I_T(\boldsymbol{\theta}_T)|}\Big(\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)^2 - \frac{\tau_T}{|I_T(\boldsymbol{\theta}_T)|}\Big(\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)^2 \\
&\quad= -\frac{15\tau_T}{16|I_T(\boldsymbol{\theta}_T)|}\Big(\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)^2. \quad \text{(S.5)}
\end{aligned}$$

Plugging (S.5) into (S.2) yields

$$0 \leq \frac{4(\gamma+1)}{3}\lambda_T\Big(\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)-\frac{2(1-2\gamma)}{3}\lambda_T\Big(\sum_{j\in I_T^c(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}|\Big)$$
$$-\frac{15\tau_T}{8|I_T(\boldsymbol{\theta}_T)|}\Big(\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)^2.$$

Rearranging the terms, we obtain

$$\sum_{j\in I_T(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}| \leq \frac{32(\gamma+1)|I_T(\boldsymbol{\theta}_T)|\lambda_T}{45\tau_T}. \quad \text{(S.6)}$$

Plugging (S.6) into (S.3) and (S.4) yields

$$\sum_{j=1}^{J_T} \bar{w}_{Tj}|\hat{\theta}_{Tj} - \theta_{Tj}| \leq \Big[\frac{32(\gamma+1)}{15(1-2\gamma)}\Big]\frac{|I_T(\boldsymbol{\theta}_T)|\lambda_T}{\tau_T} \leq \Big[\frac{16(2\gamma+5)}{3(1-2\gamma)}\Big]\frac{|I_T(\boldsymbol{\theta}_T)|\lambda_T}{\tau_T}$$
$$\text{and } \mathbb{E}_n[\Phi_T^T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T)]^2 \leq \Big[\frac{64(\gamma+1)^2}{135}\Big]\frac{|I_T(\boldsymbol{\theta}_T)|\lambda_T^2}{\tau_T} \leq \Big[\frac{16(2\gamma+5)^2}{27}\Big]\frac{|I_T(\boldsymbol{\theta}_T)|\lambda_T^2}{\tau_T},$$

on the event $\Omega_{T,1}(\boldsymbol{\theta}_T) \cap \Omega_{T,2}(\boldsymbol{\theta}_T)$. Thus, on the event $\Omega_{T,1}(\boldsymbol{\theta}_T)$, we have

$$\begin{aligned} E[\Phi_T^T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T)]^2 &\leq \frac{(1-2\gamma)^2\tau_T}{144|I_T(\boldsymbol{\theta}_T)|}\Big(\sum_{j=1}^{J_T} \bar{w}_{Tj}|\hat{\theta}_{Tj} - \theta_{Tj}|\Big)^2 + \mathbb{E}_n[\Phi_T^T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T)]^2 \\ &\leq \Big[\frac{64(2\gamma+5)^2}{81}\Big]\frac{|I_T(\boldsymbol{\theta}_T)|\lambda_T^2}{\tau_T}. \end{aligned}$$

Hence, (B.8) and (B.9) hold on the event $\Omega_{T,1}(\boldsymbol{\theta}_T) \cap \Omega_{T,2}(\boldsymbol{\theta}_T)$.

Now we prove the results for $t < T$. Assume we have

$$\sum_{j=1}^{J_s} \bar{w}_{sj}|\hat{\theta}_{sj} - \theta_{sj}| \leq \frac{16(2\gamma+5)}{3(1-2\gamma)\lambda_s} \max_{s' \in \{s,\ldots,T\}} \Big\{c_{s,s'}\frac{|I_{s'}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\Big\}, \tag{S.7}$$

$$E[\Phi_s^T(\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s)]^2 \leq \frac{64(2\gamma+5)^2}{81} \max_{s' \in \{s,\ldots,T\}} \Big\{c_{s,s'}\frac{|I_{s'}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\Big\}, \tag{S.8}$$

$$\text{and } \mathbb{E}_n[\Phi_s^T(\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s)]^2 \leq \frac{16(2\gamma+5)^2}{27} \max_{s' \in \{s,\ldots,T\}} \Big\{c_{s,s'}\frac{|I_{s'}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\Big\}, \tag{S.9}$$

where $c_{s,s} = 1$, $c_{s,s'} = 2(2\gamma+5)(5S+3)(T-s)^2 c_{s+1,s'}/9$ for $s' = s+1,\ldots,T$ and $s = t+1,\ldots,T$, on the event $\cap_{s=t+1}^T \Big\{\Omega_{s,1}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T) \cap \Omega_{s,2}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T) \cap \Omega_{s+1,3}(\boldsymbol{\theta}_{s+1},\ldots,\boldsymbol{\theta}_T)\Big\}$.

Using similar arguments as above, $\hat{\boldsymbol{\theta}}_t$ satisfies the first order condition:

$$-2\mathbb{E}_n[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1},\ldots,\hat{\boldsymbol{\theta}}_T) - \Phi_t^T\hat{\boldsymbol{\theta}}_t)\phi_{tj}] + \lambda_t w_{tj}\text{sgn}(\hat{\theta}_{tj}) = 0 \text{ for } j = 1,\ldots,J_t.$$

Thus

$$-2\mathbb{E}_n[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1},\ldots,\hat{\boldsymbol{\theta}}_T)-\Phi_t^T\hat{\boldsymbol{\theta}}_t)\Phi_t^T\boldsymbol{\theta}_t]+\lambda_t\sum_{j=1}^{J_t}w_{tj}\mathrm{sgn}(\hat{\theta}_{tj})\theta_{tj}=0$$

for any $\boldsymbol{\theta}_t\in\mathbb{R}^{J_t}$. In particular, $-2\mathbb{E}_n[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1},\ldots,\hat{\boldsymbol{\theta}}_T)-\Phi_t^T\hat{\boldsymbol{\theta}}_t)\Phi_t^T\hat{\boldsymbol{\theta}}_t]+\lambda_t\sum_{j=1}^{J_t}w_{tj}|\hat{\theta}_{tj}|=0$. Hence, for any $\boldsymbol{\theta}_t\in\mathbb{R}^{J_t}$ on the event $\Omega_{t,2}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)$, we have

$$\begin{aligned}
0&\le 2\mathbb{E}_n[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1},\ldots,\hat{\boldsymbol{\theta}}_T)-\Phi_t^T\hat{\boldsymbol{\theta}}_t)\Phi_t^T(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]+\lambda_t\sum_{j=1}^{J_t}w_{tj}|\theta_{tj}|-\lambda_t\sum_{j=1}^{J_t}w_{tj}|\hat{\theta}_{tj}| \qquad \text{(S.10)}\\
&=2\mathbb{E}_n[(\tilde{Y}_t(\boldsymbol{\theta}_{t+1},\ldots,\boldsymbol{\theta}_T)-\Phi_t^T\boldsymbol{\theta}_t)\Phi_t^T(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]+\lambda_t\sum_{j=1}^{J_t}w_{tj}|\theta_{tj}|-\lambda_t\sum_{j=1}^{J_t}w_{tj}|\hat{\theta}_{tj}|\\
&\quad+2\mathbb{E}_n[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1},\ldots,\hat{\boldsymbol{\theta}}_T)-\tilde{Y}_t(\boldsymbol{\theta}_{t+1},\ldots,\boldsymbol{\theta}_T))\Phi_t^T(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]-2\mathbb{E}_n[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]^2\\
&\le\frac{4(\gamma+1)}{3}\lambda_t\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)-\frac{2(1-2\gamma)}{3}\lambda_t\Big(\sum_{j\in I_t^c(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}|\Big)-\mathbb{E}_n[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]^2\\
&\quad+\mathbb{E}_n\big[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1},\ldots,\hat{\boldsymbol{\theta}}_T)-\tilde{Y}_t(\boldsymbol{\theta}_{t+1},\ldots,\boldsymbol{\theta}_T))\big]^2. \qquad \text{(S.11)}
\end{aligned}$$

Below, we derive an upper bound for the last term in (S.11). Note that

$$\begin{aligned}
&\mathbb{E}_n\big[\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1},\ldots,\hat{\boldsymbol{\theta}}_T)-\tilde{Y}_t(\boldsymbol{\theta}_{t+1},\ldots,\boldsymbol{\theta}_T)\big]^2\\
&=\mathbb{E}_n\Big[\sum_{s=t+1}^{T}\Big[\max_{a_s}\Phi_s^\top(H_s,a_s)\hat{\boldsymbol{\theta}}_s-\max_{a_s}\Phi_s^\top(H_s,a_s)\boldsymbol{\theta}_s-\Phi_s^\top(H_s,A_s)(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\Big]\Big]^2\\
&\le 2(T-t)\sum_{s=t+1}^{T}\Big\{\mathbb{E}_n\Big[\max_{a_s}\big|\Phi_s^\top(H_s,a_s)(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\big|^2\Big]+\mathbb{E}_n\Big[\Phi_s^\top(H_s,A_s)(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\Big]^2\Big\}. \qquad \text{(S.12)}
\end{aligned}$$

We can further show that

$$\begin{aligned}
&\mathbb{E}_n\Big[\max_{a_s}\big|\Phi_s^\top(H_s,a_s)(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\big|^2\Big]\\
&\le(\mathbb{E}_n-E)\Big[\textstyle\sum_{a_s}\big|\Phi_s^\top(H_s,a_s)(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\big|^2\Big]+E\Big[\textstyle\sum_{a_s}\big|\Phi_s^\top(H_s,a_s)(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\big|^2\Big]\\
&\le\max_{j,k\in\{1,\ldots,J_s\}}\Big|(\mathbb{E}_n-E)\Big(\sum_{a_s\in\mathscr{A}_s}\frac{\phi_{sj}(H_s,a_s)\phi_{sk}(H_s,a_s)}{\bar{w}_{sj}\bar{w}_{sk}}\Big)\Big|\Big(\sum_{j=1}^{J_s}\bar{w}_{sj}|\hat{\theta}_{sj}-\theta_{sj}|\Big)^2
\end{aligned}$$

$$
\begin{aligned}
&\qquad + E\Big[\sum_{a_s\in\mathscr{A}_s} p_s(a_s|H_s)S\big[\Phi_s^\top(H_s,a_s)(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\big]^2\Big] \\
&\le \frac{(1-2\gamma)^2|\mathscr{A}_s|}{144\max_{s'\in\{s,\ldots,T\}}\{|I_{s'}(\boldsymbol{\theta}_{s'})|/\tau_{s'}\}}\left(\frac{16(2\gamma+5)}{3(1-2\gamma)\lambda_s}\max_{s'\in\{s,\ldots,T\}}\left\{c_{s,s'}\frac{|I_{s'}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\right\}\right)^2 \\
&\qquad + SE\big[\Phi_s^\top(H_s,A_s)(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\big]^2 \\
&\le \frac{16(2\gamma+5)^2|\mathscr{A}_s|}{81}\left[\max_{s'\in\{s,\ldots,T\}}\left\{c_{s,s'}\frac{|I_{s'}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\right\}\right] \\
&\qquad + \frac{64(2\gamma+5)^2S}{81}\left[\max_{s'\in\{s,\ldots,T\}}\left\{c_{s,s'}\frac{|I_{s'}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\right\}\right] \\
&\le \frac{80(2\gamma+5)^2S}{81}\left[\max_{s'\in\{s,\ldots,T\}}\left\{c_{s,s'}\frac{|I_{s'}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\right\}\right] \qquad \text{(S.13)}
\end{aligned}
$$

where the second inequality follows from the assumption that $p_s(a_s|h_s) \ge S^{-1}$ for all (h_s, a_s) pairs, the third inequality follows from the definition of $\Omega_{s,3}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T)$ and (S.7), the fourth inequality follows from (S.8) and the assumption that $\lambda_s^2 \ge c_{s,s'}\lambda_{s'}^2$ for $s \le s'$, and the last inequality follows from the fact that $|\mathscr{A}_s| \le S$. Plugging 3 (S.9) and (S.13) into (S.12) and noticing that $c_{s,s'} \le c_{t+1,s'}$ for any $s \ge t+1$ and $s' \ge s$, we have

$$
\mathbb{E}_n\big[\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1},\ldots,\hat{\boldsymbol{\theta}}_T)-\tilde{Y}_t(\boldsymbol{\theta}_{t+1},\ldots,\boldsymbol{\theta}_T)\big]^2 \le C_t\max_{s\in\{t+1,\ldots,T\}}\left\{c_{t+1,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}, \qquad \text{(S.14)}
$$

where $C_t = 32(2\gamma+5)^2(5S+3)(T-t)^2/81$. This, together with (S.11), implies that, on the event $\cap_{s=t+1}^T\Big\{\Omega_{s,1}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T)\cap\Omega_{s,2}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T)\cap\Omega_{s,3}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T)\Big\}\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)$,

$$
\begin{aligned}
0 \le{}& \frac{4(\gamma+1)}{3}\lambda_t\sum_{j\in I_t(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}| - \frac{2(1-2\gamma)}{3}\lambda_t\sum_{j\in I_t^c(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}| - \mathbb{E}_n[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]^2 \\
&+ C_t\max_{s\in\{t+1,\ldots,T\}}\left\{c_{t+1,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}. \qquad \text{(S.15)}
\end{aligned}
$$

Thus

$$\sum_{j\in I_t^c(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\hat{\theta}_{tj}| \le \frac{2(\gamma+1)}{1-2\gamma}\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big) + \frac{3C_t}{2(1-2\gamma)\lambda_t}\max_{s\in\{t+1,\ldots,T\}}\left\{c_{t+1,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}$$
$$\text{and } \mathbb{E}_n[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]^2 \le \frac{4(\gamma+1)}{3}\lambda_t\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big) + C_t\max_{s\in\{t+1,\ldots,T\}}\left\{c_{t+1,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}. \tag{S.16}$$

If $I_t(\boldsymbol{\theta}_t)$ is empty (i.e., $\boldsymbol{\theta}_t = \mathbf{0}$), then it is easy to verify that (B.8) and (B.9) hold. If $I_t(\boldsymbol{\theta}_t)$ is non-empty, define the sets

$$\Theta_{t,1}(\boldsymbol{\theta}_t) = \Bigg\{\tilde{\boldsymbol{\theta}}_t \in \mathbb{R}^{J_t} : \sum_{j\in I_t^c(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj}| \le \frac{2(\gamma+1)}{1-2\gamma}\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj}-\theta_{tj}|\Big) + \frac{3C_t}{2(1-2\gamma)\lambda_t}\max_{s\in\{t+1,\ldots,T\}}\left\{c_{t+1,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}\Bigg\},$$

$$\Theta_{t,2}(\boldsymbol{\theta}_t) = \Bigg\{\tilde{\boldsymbol{\theta}}_t \in \mathbb{R}^{J_t} : \sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj}-\theta_{tj}| > \max\left\{\frac{8(2\gamma+5)|I_t(\boldsymbol{\theta}_t)|\lambda_t}{9\tau_t}, \frac{C_t}{2\lambda_t}\max_{s\in\{t+1,\ldots,T\}}\left\{c_{t+1,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}\right\}\Bigg\}.$$

On the event $\cap_{s=t}^T\Big\{\Omega_{s,1}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T)\cap\Omega_{s,2}(\boldsymbol{\theta}_s,\ldots,\boldsymbol{\theta}_T)\cap\Omega_{s+1,3}(\boldsymbol{\theta}_{s+1},\ldots,\boldsymbol{\theta}_T)\Big\}$, we have $\hat{\boldsymbol{\theta}}_t \in \Theta_{t,1}(\boldsymbol{\theta}_t)$. Thus, $\hat{\boldsymbol{\theta}}_t \in \Theta_{t,1}(\boldsymbol{\theta}_t)$ on the event $\Omega_{t,2}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)$. Note that condition (3.8) by Assumption (B3) implies that

$$E[\Phi_t^T(\tilde{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)]^2 \ge \frac{\tau_t(\sum_{j\in I_t(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\tilde{\theta}_{tj}-\theta_{tj}|)^2}{|I_t(\boldsymbol{\theta}_t)|} \tag{S.17}$$

for any $\boldsymbol{\theta}_t \in \Theta_t$ and $\tilde{\boldsymbol{\theta}}_t \in \Theta_{t,1}(\boldsymbol{\theta}_t)$. In addition,

$$\sup_{\tilde{\boldsymbol{\theta}}_t\in\Theta_{t,1}(\boldsymbol{\theta}_t)\cap\Theta_{t,2}(\boldsymbol{\theta}_t)}\left\{\frac{4(\gamma+1)}{3}\lambda_t\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\tilde{\theta}_{tj}-\theta_{tj}|\Big) - \frac{2(1-2\gamma)}{3}\lambda_t\Big(\sum_{j\in I_t^c(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\tilde{\theta}_{tj}|\Big)\right.$$

$$
\begin{aligned}
&\qquad - \mathbb{E}_n[\Phi_t^\top(\tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 + C_t \max_{s\in\{t+1,\ldots,T\}} \left\{ c_{t+1,s} \frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \Bigg\} \\
\leq & \sup_{\tilde{\boldsymbol{\theta}}_t \in \Theta_{t,1}(\boldsymbol{\theta}_t)\cap\Theta_{t,2}(\boldsymbol{\theta}_t)} \Bigg\{ \frac{4(\gamma+1)}{3}\lambda_t\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big) - \frac{2(1-2\gamma)}{3}\lambda_t\Big(\sum_{j\in I_t^c(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj}| \Big) \\
&\qquad + \max_{j,k\in\{1,\ldots,J_t\}} \Big| (E - \mathbb{E}_n)\Big(\frac{\phi_{tj}\phi_{tk}}{\bar{w}_{tj}\bar{w}_{tk}} \Big) \Big| \Big(\sum_{j=1}^{J_t} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big)^2 \\
&\qquad - E[\Phi_t^\top(\tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 + C_t \max_{s\in\{t+1,\ldots,T\}} \left\{ c_{t+1,s} \frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \Bigg\} \\
\leq & \sup_{\tilde{\boldsymbol{\theta}}_t \in \Theta_{t,1}(\boldsymbol{\theta}_t)\cap\Theta_{t,2}(\boldsymbol{\theta}_t)} \Bigg\{ \frac{4(\gamma+1)}{3}\lambda_t\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big) \\
&\quad + \frac{(1-2\gamma)^2}{144\max_s\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}} \Bigg[\frac{3}{1-2\gamma} \Bigg(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| + \frac{C_t}{2\lambda_t} \max_{s\in\{t+1,\ldots,T\}} \left\{ c_{t+1,s} \frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \Bigg) \Bigg]^2 \\
&\qquad - E[\Phi_t^\top(\tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t]^2 + C_t \max_{s\in\{t+1,\ldots,T\}} \left\{ c_{t+1,s} \frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \right\} \Bigg\} \\
\leq & \sup_{\tilde{\boldsymbol{\theta}}_t \in \Theta_{t,1}(\boldsymbol{\theta}_t)\cap\Theta_{t,2}(\boldsymbol{\theta}_t)} \Bigg\{ \frac{4(\gamma+1)}{3}\lambda_t\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big) + \frac{\tau_t}{4|I_t(\boldsymbol{\theta}_t)|}\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big)^2 \\
&\qquad - \frac{\tau_t}{|I_t(\boldsymbol{\theta}_t)|}\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big)^2 + 2\lambda_t\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big) \Bigg\} \\
\leq & \sup_{\tilde{\boldsymbol{\theta}}_t \in \Theta_{t,1}(\boldsymbol{\theta}_t)\cap\Theta_{t,2}(\boldsymbol{\theta}_t)} \Bigg\{ \Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big) \times \\
&\qquad \Big[\frac{2(2\gamma+5)}{3}\lambda_t - \frac{3\tau_t}{4|I_t(\boldsymbol{\theta}_t)|}\Big(\sum_{j\in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| \Big) \Big] \Bigg\} \\
< & \, 0,
\end{aligned}
$$

where the second inequality follows from the definition of $\Omega_{t,1}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)$ and $\Theta_{t,1}(\boldsymbol{\theta}_t)$, the third inequality follows from the definition of $\Theta_{t,2}(\boldsymbol{\theta}_t)$ and (S.17), and the last inequality follows from the definition of $\Theta_{t,2}(\boldsymbol{\theta}_t)$.

Since $\hat{\boldsymbol{\theta}}_t$ satisfies inequality (S.15), we have $\hat{\boldsymbol{\theta}}_t \in \Theta_{t,1}(\boldsymbol{\theta}_t)\cap\Theta_{t,2}(\boldsymbol{\theta}_t)^C$ on the event $\cap_{s=t}^T \Big\{ \Omega_{s,1}(\boldsymbol{\theta}_s,$

$\ldots, \boldsymbol{\theta}_T) \cap \Omega_{s,2}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s+1,3}(\boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_T)\Big\}$. This, together with (S.16), implies that

$$\begin{aligned}
\sum_{j=1}^{J_t} \bar{w}_{tj}|\hat{\theta}_{tj} - \theta_{tj}| &\le \max\left\{\frac{16(2\gamma+5)|I_t(\boldsymbol{\theta}_t)|\lambda_t}{3(1-2\gamma)\tau_t}, \frac{3C_t}{(1-2\gamma)\lambda_t} \max_{s\in\{t+1,\ldots,T\}}\left\{c_{t+1,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}\right\} \\
&= \frac{16(2\gamma+5)}{3(1-2\gamma)\lambda_t} \max_{s\in\{t,\ldots,T\}}\left\{c_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\} \\
\text{and } \mathbb{E}_n[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 &\le \max\left\{\frac{16(2\gamma+5)^2|I_t(\boldsymbol{\theta}_t)|\lambda_t^2}{27\tau_t}, \frac{(2\gamma+5)C_t}{3} \max_{s\in\{t+1,\ldots,T\}}\left\{c_{t+1,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}\right\} \\
&= \frac{16(2\gamma+5)^2}{27} \max_{s\in\{t,\ldots,T\}}\left\{c_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}.
\end{aligned}$$

where $c_{t,t} = 1, c_{t,s} = 9C_t c_{t+1,s}/[16(2\gamma+5)] = 2(2\gamma+5)(5S+3)(T-t)^2 c_{t+1,s}/9$. In addition, we have

$$\begin{aligned}
E[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 &\le \frac{(1-2\gamma)^2}{144\max_{s\in\{t,\ldots,T\}}\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}}\Big(\sum_{j=1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj} - \theta_{tj}|\Big)^2 + \mathbb{E}_n[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 \\
&\le \frac{64(2\gamma+5)^2}{81} \max_{s\in\{t,\ldots,T\}}\left\{c_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\},
\end{aligned}$$

where the last inequality follows from the condition that $\lambda_t^2 \ge c_{t,s}\lambda_s^2$ for $t \le s$. This completes the proof. □

Proof of Lemma 5.

Similarly as in the proof of Lemma 4, we prove the results using induction. Consider fixed n and fixed $\boldsymbol{\theta}_T \in \Theta_T$. Since $E[\Phi_{T2}^\top(H_T, A_T)|H_T] = \mathbf{0}$ a.s., we have $E(\Phi_{T1}\Phi_{T2}^T) = \mathbf{0}_{J_{T1}\times J_{T2}}$. On the event $\Omega_{T,1}(\boldsymbol{\theta}_T)$, we have

$$\begin{aligned}
&\mathbb{E}_n\big[\Phi_T^T(\boldsymbol{\theta}_T - \hat{\boldsymbol{\theta}}_T)\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2} - \boldsymbol{\theta}_{T2})\big] \\
&\quad = (E - \mathbb{E}_n)[\Phi_T^T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T)\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2} - \boldsymbol{\theta}_{T2})] - E[\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2} - \boldsymbol{\theta}_{T2})]^2
\end{aligned}$$

$$
\begin{aligned}
&\le \max_{j,k\in\{1,\ldots,J_T\}} \Big|(E-\mathbb{E}_n)\Big(\frac{\phi_{Tj}\phi_{Tk}}{\bar{w}_{Tj}\bar{w}_{Tk}}\Big)\Big|\Big(\sum_{j=1}^{J_T}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)\Big(\sum_{j=J_{T1}+1}^{J_T}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)\\
&\quad - E[\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})]^2\\
&\le \frac{(1-2\gamma)(2\gamma+5)}{27}\lambda_T\Big(\sum_{j=J_{T1}+1}^{J_T}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big) - E[\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})]^2,
\end{aligned}
$$

where the last inequality follows from the definition of $\Omega_{T,1}(\boldsymbol{\theta}_T)$ and (B.8). Note that (S.1) holds for any $\boldsymbol{\theta}_T \in \mathbb{R}^{J_T}$. In particular, with $(\hat{\boldsymbol{\theta}}_{T1}^\top, \boldsymbol{\theta}_{T2}^\top)^\top$, on the event $\Omega_{T,1}(\boldsymbol{\theta}_T)\cap\Omega_{T,2}(\boldsymbol{\theta}_T)$, we have

$$
\begin{aligned}
0 &\le 2\mathbb{E}_n[(Y_T-\Phi_T^T\hat{\boldsymbol{\theta}}_T)\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})] + \lambda_T\sum_{j=J_{T1}+1}^{J_T} w_{Tj}|\theta_{Tj}| - \lambda_T\sum_{j=J_{T1}+1}^{J_T} w_{Tj}|\hat{\theta}_{Tj}|\\
&\le \frac{4\gamma+1}{3}\lambda_T\Big(\sum_{j=J_{T1}+1}^{J_T}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big) + \lambda_T\sum_{j=J_{T1}+1}^{J_T} w_{Tj}|\theta_{Tj}| - \lambda_T\sum_{j=J_{T1}+1}^{J_T} w_{Tj}|\hat{\theta}_{Tj}|\\
&\quad + \frac{2(1-2\gamma)(2\gamma+5)}{27}\lambda_T\Big(\sum_{j=J_{T1}+1}^{J_T}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big) - 2E[\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})]^2\\
&\le \frac{4(2\gamma+3)(4-\gamma)}{27}\lambda_T\Big(\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)\\
&\quad - \frac{4(1-2\gamma)(2-\gamma)}{27}\lambda_T\Big(\sum_{j\in I_{T2}^c(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}|\Big) - 2E[\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})]^2. \qquad \text{(S.18)}
\end{aligned}
$$

This implies

$$
\sum_{j\in I_{T2}^c(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}| \le \frac{(2\gamma+3)(4-\gamma)}{(1-2\gamma)(2-\gamma)}\Big(\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big) \qquad \text{(S.19)}
$$

$$
\text{and } E[\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})]^2 \le \frac{2(2\gamma+3)(4-\gamma)}{27}\lambda_T\Big(\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big). \qquad \text{(S.20)}
$$

Note that Assumption (B3) implies that

$$
E[\Phi_{T2}^T(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})]^2 \ge \frac{\tau_T(\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|)^2}{|I_{T2}(\boldsymbol{\theta}_T)|}. \qquad \text{(S.21)}
$$

Plugging (S.21) into (S.18) yields

$$
\begin{aligned}
0 \le{}& \frac{4(2\gamma+3)(4-\gamma)}{27}\lambda_T\Big(\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big) \\
&- \frac{4(1-2\gamma)(2-\gamma)}{27}\lambda_T\Big(\sum_{j\in I^c_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}|\Big) - \frac{2\tau_T(\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|)^2}{|I_{T2}(\boldsymbol{\theta}_T)|} \\
\le{}& \Big(\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)\Big[\frac{4(2\gamma+3)(4-\gamma)}{27}\lambda_T - \frac{2\tau_T}{|I_{T2}(\boldsymbol{\theta}_T)|}\Big(\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}|\Big)\Big].
\end{aligned}
$$

Thus

$$
\sum_{j\in I_{T2}(\boldsymbol{\theta}_T)}\bar{w}_{Tj}|\hat{\theta}_{Tj}-\theta_{Tj}| \le \frac{2(2\gamma+3)(4-\gamma)|I_{T2}(\boldsymbol{\theta}_T)|\lambda_T}{27\tau_T}.
$$

This, together with (S.19) and (S.20), implies that

$$
\begin{aligned}
\sum_{j=J_{T1}+1}^{J_T}\bar{w}_{Tj}|\hat{\theta}_{Tj}| &\le \Big[\frac{28(2\gamma+3)(4-\gamma)}{27(1-2\gamma)(2-\gamma)}\Big]\frac{|I_{T2}(\boldsymbol{\theta}_T)|\lambda_T}{\tau_T} \\
\text{and } E[\Phi^T_{T2}(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})]^2 &\le \Big[\frac{2(2\gamma+3)(4-\gamma)}{27}\Big]^2\frac{|I_{T2}(\boldsymbol{\theta}_T)|\lambda_T^2}{\tau_T}.
\end{aligned}
$$

Algebra suffices to show (B.10) and (B.11) for $t = T$. This also implies that

$$
\begin{aligned}
&\mathbb{E}_n[\Phi^T_{T2}(\hat{\boldsymbol{\theta}}_{T2}-\boldsymbol{\theta}_{T2})]^2 \\
&\le \max_{j,k\in\{1,\ldots,J_t\}}\Big|(\mathbb{E}_n-E)\Big(\frac{\phi_{tj}\phi_{tk}}{\bar{w}_{tj}\bar{w}_{tk}}\Big)\Big|\Big(\sum_{j=J_{t1}+1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)^2 + E[\Phi^T_{t2}(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})]^2 \\
&\le \frac{(1-2\gamma)^2\tau_T}{144|I_T(\boldsymbol{\theta}_T)|}\Big(\Big[\frac{28(2\gamma+3)(4-\gamma)}{27(1-2\gamma)(2-\gamma)}\Big]\frac{|I_{T2}(\boldsymbol{\theta}_T)|\lambda_T}{\tau_T}\Big)^2 + \Big[\frac{2(2\gamma+3)(4-\gamma)}{27}\Big]^2\frac{|I_{T2}(\boldsymbol{\theta}_T)|\lambda_T^2}{\tau_T} \\
&\le 2\Big[\frac{2(2\gamma+3)(4-\gamma)}{27}\Big]^2\frac{|I_{T2}(\boldsymbol{\theta}_T)|\lambda_T^2}{\tau_T}.
\end{aligned}
$$

Now we prove the results for $t < T$. Assume for any given $\{\boldsymbol{\theta}_{s'} \in \Theta_{s'} : s' = s, \ldots, T\}$, we

have

$$\sum_{j=J_s1+1}^{J_s} \bar{w}_{sj}|\hat{\theta}_{sj} - \theta_{sj}| \le \left[\frac{81}{(1-2\gamma)^2} - 3\right]\lambda_s^{-1} \max_{s'\in\{s,\ldots,T\}} \left\{\bar{c}_{s,s'}\frac{|I_{s'2}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\right\}$$

$$E[\Phi_{s2}^T(\hat{\boldsymbol{\theta}}_{s2} - \boldsymbol{\theta}_{s2})]^2 \le \left[3 - \frac{(1-2\gamma)^2}{9}\right]^2 \max_{s'\in\{s,\ldots,T\}} \left\{\bar{c}_{s,s'}\frac{|I_{s'2}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\right\}$$

and

$$\mathbb{E}_n[\Phi_{s2}^T(\hat{\boldsymbol{\theta}}_{s2} - \boldsymbol{\theta}_{s2})]^2 \le \left[\frac{81\max_{s'\in\{s,\ldots,T\}}\{\bar{c}_{s,s'}/c_{s,s'}\}}{16(1-2\gamma)^2} + 1\right]\left[3 - \frac{(1-2\gamma)^2}{9}\right]^2 \max_{s'\in\{s,\ldots,T\}} \left\{\bar{c}_{s,s'}\frac{|I_{s'2}(\boldsymbol{\theta}_{s'})|\lambda_{s'}^2}{\tau_{s'}}\right\},$$

for $s' = s{+}1, \ldots, T$ and $s = T, \ldots, t{+}1$, on the event $\cap_{s=t+1}^T \Big\{\Omega_{s,1}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s,2}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s+1,3}(\boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_T)\Big\}$.

Note that (S.10) holds for any $\boldsymbol{\theta}_t \in \mathbb{R}^{J_t}$. In particular, with $\boldsymbol{\theta}_t = (\hat{\boldsymbol{\theta}}_{t1}^\top, \boldsymbol{\theta}_{t2}^\top)^\top$, we have on the event $\cap_{s=t}^T \big\{\Omega_{s,1}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s,2}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s+1,3}(\boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_T)\big\}$,

$$\begin{aligned}
0 &\le 2\mathbb{E}_n[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1}, \ldots, \hat{\boldsymbol{\theta}}_T) - \Phi_t^T\hat{\boldsymbol{\theta}}_t)\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})] + \lambda_t \sum_{j=J_{t1}+1}^{J_t} w_{tj}|\theta_{tj}| - \lambda_t \sum_{j=J_{t1}+1}^{J_t} w_{tj}|\hat{\theta}_{tj}| \\
&\le 2\mathbb{E}_n[(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) - \Phi_t^T\boldsymbol{\theta}_t)\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})] + \lambda_t \sum_{j=J_{t1}+1}^{J_t} w_{tj}|\theta_{tj}| - \lambda_t \sum_{j=J_{t1}+1}^{J_t} w_{tj}|\hat{\theta}_{tj}| \\
&\quad + 2\mathbb{E}_n[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1}, \ldots, \hat{\boldsymbol{\theta}}_T) - \tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T))\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})] \\
&\qquad\qquad - 2\mathbb{E}_n[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)\Phi_{t2}^\top(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})] \\
&\le 2\mathbb{E}_n[(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) - \Phi_t^T\boldsymbol{\theta}_t)\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})] + \lambda_t \sum_{j=J_{t1}+1}^{J_t} w_{tj}|\theta_{tj}| - \lambda_t \sum_{j=J_{t1}+1}^{J_t} w_{tj}|\hat{\theta}_{tj}| \\
&\quad - E[\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})]^2 + \mathbb{E}_n[\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1}, \ldots, \hat{\boldsymbol{\theta}}_T) - \tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T)]^2 \\
&\quad + \{(\mathbb{E}_n - E)[\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})]^2 + 2(E - \mathbb{E}_n)[\Phi_t^T(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})]\}, \qquad \text{(S.22)}
\end{aligned}$$

where the last inequality follows from AM-GM inequality and the condition that $E(\Phi_{t1}\Phi_{t2}^\top) = \mathbf{0}$. Below, we derive upper bounds for the last two terms of (S.22). Since $\Phi_{s1}(H_s)$ does not involve treatment A_s, it is easy to see that $Y_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T)$ defined in (B.7) can be re-written as

$$\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) = Y_t + \sum_{s=t+1}^{T} \Big[Y_s + \max_{a_s} \Phi_{s2}^T(H_s, a_s)\boldsymbol{\theta}_{s2} - \Phi_{s2}^T(H_s, A_s)\boldsymbol{\theta}_{s2} \Big].$$

Note that on event $\Omega_{s,3}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T)$

$$\begin{aligned}
&\max_{j,k\in\{1,\ldots,J_t\}} \Big| (E - \mathbb{E}_n)\Big(\sum_{a_s \in \mathscr{A}_s} \frac{\phi_{sj}(H_s, a_s)\phi_{sk}(H_s, a_s)}{\bar{w}_{sj}\bar{w}_{sk}} \Big) \Big| \\
&\qquad \le \frac{(1-2\gamma)^2 |\mathscr{A}_s| \max_{s\in\{t,\ldots,T\}} \{\bar{c}_{t,s}/c_{t,s}\}\lambda_t^2}{144 \max_{s\in\{t,\ldots,T\}} \big\{ \big[\max_{s\in\{t,\ldots,T\}} \{\bar{c}_{t,s}\lambda_t^2/c_{t,s}\} \big] I_{s2}(\boldsymbol{\theta}_s)/\tau_s \big\}} \\
&\qquad \le \frac{(1-2\gamma)^2 |\mathscr{A}_s| \max_{s'\in\{s,\ldots,T\}} \{\bar{c}_{s,s'}/c_{s,s'}\}\lambda_s^2}{144 \max_{s'\in\{s,\ldots,T\}} \big\{ \bar{c}_{s,s'} I_{s'2}(\boldsymbol{\theta}_{s'})\lambda_{s'}^2/\tau_{s'} \big\}}.
\end{aligned}$$

Using the similar arguments as in the proof of Lemma 4, we can show that

$$\begin{aligned}
&\mathbb{E}_n[(\tilde{Y}_t(\hat{\boldsymbol{\theta}}_{t+1}, \ldots, \hat{\boldsymbol{\theta}}_T) - \tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T)]^2 \\
&= \mathbb{E}_n \Big[\sum_{s=t+1}^{T} \Big[\max_{a_s} \Phi_{s2}^\top(H_s, a_s)\hat{\boldsymbol{\theta}}_{s2} - \max_{a_s} \Phi_{s2}^\top(H_s, a_s)\boldsymbol{\theta}_{s2} - \Phi_{s2}^\top(H_s, A_s)(\hat{\boldsymbol{\theta}}_{s2} - \boldsymbol{\theta}_{s2}) \Big] \Big]^2 \\
&\le 2(T-t) \sum_{s=t+1}^{T} \Big\{ \mathbb{E}_n \Big[\max_{a_s} \big| \Phi_{s2}^\top(H_s, a_s)(\hat{\boldsymbol{\theta}}_{s2} - \boldsymbol{\theta}_{s2}) \big|^2 \Big] + \mathbb{E}_n \Big[\Phi_{s2}^\top(H_s, A_s)(\hat{\boldsymbol{\theta}}_{s2} - \boldsymbol{\theta}_{s2}) \Big]^2 \Big\} \\
&\le C_{t2} \max_{s\in\{t+1,\ldots,T\}} \Big\{ \bar{c}_{t+1,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s} \Big\} \qquad \text{(S.23)}
\end{aligned}$$

where $C_{t2} = 2(T-t)^2(S+1)\Big[\frac{81 \max_{s\in\{t+1,\ldots,T\}}\{\bar{c}_{t+1,s}/c_{t+1,s}\}}{16(1-2\gamma)^2} + 1 \Big] \Big[3 - \frac{(1-2\gamma)^2}{9} \Big]^2$. In addition,

$$\begin{aligned}
&(\mathbb{E}_n - E)[\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})]^2 + 2(E - \mathbb{E}_n)[\Phi_t^T(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})] \\
&\qquad \le 3 \max_{j,k\in\{1,\ldots,J_t\}} \Big| (E - \mathbb{E}_n)\Big(\frac{\phi_{tj}\phi_{tk}}{\bar{w}_{tj}\bar{w}_{tk}} \Big) \Big| \Big(\sum_{j=1}^{J_t} \bar{w}_{tj}|\hat{\theta}_{tj} - \theta_{tj}| \Big) \Big(\sum_{j=J_{t1}+1}^{J_t} \bar{w}_{tj}|\hat{\theta}_{tj} - \theta_{tj}| \Big)
\end{aligned}$$

$$\leq \frac{(1-2\gamma)(2\gamma+5)}{9}\lambda_t\Big(\sum_{j=J_{t1}+1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big), \tag{S.24}$$

where the last inequality follows from the definition of $\Omega_{t,1}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)$, (B.8), and the assumption that $\lambda_t^2 \geq c_{t,s}\lambda_s^2$ for $t \leq s$. Plugging in (S.23) and (S.24) into (S.22) yields

$$\begin{aligned}
0 \leq{}& \frac{4\gamma+1}{3}\lambda_t\Big(\sum_{j=J_{t1}+1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)+\lambda_t\sum_{j=J_{t1}+1}^{J_t}w_{tj}|\theta_{tj}|-\lambda_t\sum_{j=J_{t1}+1}^{J_t}w_{tj}|\hat{\theta}_{tj}|-E[\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})]^2\\
&+C_{t2}\max_{s\in\{t+1,\ldots,T\}}\Big\{\bar{c}_{t+1,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\Big\}+\frac{(1-2\gamma)(2\gamma+5)}{9}\lambda_t\Big(\sum_{j=J_{t1}+1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)\\
\leq{}&\Big[2-\frac{(1-2\gamma)^2}{9}\Big]\lambda_t\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)-\frac{(1-2\gamma)^2}{9}\lambda_t\Big(\sum_{j\in I_{t2}^c(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}|\Big)\\
&-E[\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})]^2+C_{t2}\max_{s\in\{t+1,\ldots,T\}}\Big\{\bar{c}_{t+1,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\Big\}.
\end{aligned} \tag{S.25}$$

This implies

$$\begin{aligned}
\sum_{j\in I_{t2}^c(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}| \leq{}& \Big[\frac{18}{(1-2\gamma)^2}-1\Big]\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)\\
&+\frac{9C_{t2}}{(1-2\gamma)^2\lambda_t}\max_{s\in\{t+1,\ldots,T\}}\Big\{\bar{c}_{t+1,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\Big\},
\end{aligned}$$

and

$$\begin{aligned}
E[\Phi_{t2}^\top(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})]^2 \leq{}& \Big[2-\frac{(1-2\gamma)^2}{9}\Big]\lambda_t\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)\\
&+C_{t2}\max_{s\in\{t+1,\ldots,T\}}\Big\{\bar{c}_{t+1,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\Big\}.
\end{aligned}$$

If $I_{t2}(\boldsymbol{\theta}_t)$ is empty (i.e., $\boldsymbol{\theta}_{t2}=\mathbf{0}$), then (B.10) and (B.11) hold. If $I_{t2}(\boldsymbol{\theta}_t)$ is non-empty, define

the sets

$$\acute{\Theta}_1(\boldsymbol{\theta}_t) = \Big\{\tilde{\boldsymbol{\theta}}_t \in \mathbb{R}^{J_{t2}} : \sum_{j\in I_{t2}^c(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj}| \le \Big[\frac{18}{(1-2\gamma)^2} - 1\Big]\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}|\Big) + \frac{9C_{t2}}{(1-2\gamma)^2\lambda_t} \max_{s\in\{t+1,\ldots,T\}}\Big\{\bar{c}_{t+1,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\Big\}\Big\},$$

$$\acute{\Theta}_2(\boldsymbol{\theta}_t) = \Big\{\tilde{\boldsymbol{\theta}}_t \in \mathbb{R}^{J_{t2}} : \sum_{j\in I_{t2}(\boldsymbol{\theta}_t)} w_{tj}|\tilde{\theta}_{tj} - \theta_{tj}| > \max\Big\{\Big[3 - \frac{(1-2\gamma)^2}{9}\Big]\frac{|I_{t2}(\boldsymbol{\theta}_t)|\lambda_t}{\tau_t}, \frac{C_{t2}}{\lambda_t}\max_{s\in\{t+1,\ldots,T\}}\Big\{\bar{c}_{t+1,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\Big\}\Big\}\Big\}.$$

Thus, $\hat{\boldsymbol{\theta}}_{t2} \in \acute{\Theta}_1(\boldsymbol{\theta}_t)$ on the event $\Omega_{t,2}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)$. Next, for any $\boldsymbol{\theta}_{t2} \in \Theta_t$ and $\tilde{\boldsymbol{\theta}}_{t2} \in \acute{\Theta}_1(\boldsymbol{\theta}_t)$, note that condition (3.8) by Assumption (B3) implies that

$$E[\Phi_{t2}^T(\tilde{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})]^2 \ge \frac{\tau_t(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}|)^2}{|I_{t2}(\boldsymbol{\theta}_t)|}. \tag{S.26}$$

In addition, on the event $\cap_{s=t}^T\big\{\Omega_{s,1}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s,2}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s+1,3}(\boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_T)\big\}$,

$$\begin{aligned}
&\sup_{\tilde{\boldsymbol{\theta}}_t\in\acute{\Theta}_1(\boldsymbol{\theta}_t)\cap\acute{\Theta}_2(\boldsymbol{\theta}_t)} \Big\{\Big[2 - \frac{(1-2\gamma)^2}{9}\Big]\lambda_t\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}|\Big) - \frac{(1-2\gamma)^2}{9}\lambda_t\Big(\sum_{j\in I_{t2}^c(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj}|\Big) \\
&\qquad - E[\Phi_{t2}^T(\tilde{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})]^2 + C_{t2}\max_{s\in\{t+1,\ldots,T\}}\Big\{\bar{c}_{t+1,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\Big\}\Big\} \\
\le\ &\sup_{\tilde{\boldsymbol{\theta}}_t\in\acute{\Theta}_1(\boldsymbol{\theta}_t)\cap\acute{\Theta}_2(\boldsymbol{\theta}_t)} \Big\{\Big[2 - \frac{(1-2\gamma)^2}{9}\Big]\lambda_t\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}|\Big) \\
&\qquad - \frac{\tau_t}{|I_{t2}(\boldsymbol{\theta}_t)|}\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}|\Big)^2 + C_{t2}\max_{s\in\{t+1,\ldots,T\}}\Big\{\bar{c}_{t+1,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\Big\}\Big\} \\
\le\ &\sup_{\tilde{\boldsymbol{\theta}}_t\in\acute{\Theta}_1(\boldsymbol{\theta}_t)\cap\acute{\Theta}_2(\boldsymbol{\theta}_t)} \Big\{\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}|\Big)\times \\
&\qquad \Big[\Big[3 - \frac{(1-2\gamma)^2}{9}\Big]\lambda_t - \frac{\tau_t}{|I_{t2}(\boldsymbol{\theta}_t)|}\Big(\sum_{j\in I_{t2}(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\tilde{\theta}_{tj} - \theta_{tj}|\Big)\Big]\Big\} \\
<\ &0,
\end{aligned}$$

where the first inequality follows from (S.26), and the last two inequalities follow from the definition of $\acute{\Theta}_2(\boldsymbol{\theta}_t)$.

Since $\hat{\boldsymbol{\theta}}_{t2}$ satisfies inequality (S.25), we have $\hat{\boldsymbol{\theta}}_{t2} \in \acute{\Theta}_1(\boldsymbol{\theta}_t) \cap \acute{\Theta}_2(\boldsymbol{\theta}_t)^C$ on the event $\cap_{s=t}^{T} \{\Omega_{s,1}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s,2}(\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_T) \cap \Omega_{s+1,3}(\boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_T)\}$. Algebra suffices to show

$$\begin{aligned}
&\sum_{j=J_{t1}+1}^{J_s} \bar{w}_{tj} |\hat{\theta}_{tj} - \theta_{tj}| \\
&\quad \le \max\left\{ \left[\frac{81}{(1-2\gamma)^2} - 3 \right] \frac{|I_{t2}(\boldsymbol{\theta}_t)| \lambda_t}{\tau_t}, \frac{27 C_{t2}}{(1-2\gamma)^2 \lambda_t} \max_{s \in \{t+1,\ldots,T\}} \left\{ \bar{c}_{t+1,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)| \lambda_s^2}{\tau_s} \right\} \right\} \\
&\quad = \left[\frac{81}{(1-2\gamma)^2} - 3 \right] \lambda_t^{-1} \max_{s \in \{t,\ldots,T\}} \left\{ \bar{c}_{t,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)| \lambda_s^2}{\tau_s}, \right\}
\end{aligned}$$

and

$$\begin{aligned}
&E\left[\Phi_{t2}^T (\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})\right]^2 \\
&\quad \le \max\left\{ \left[3 - \frac{(1-2\gamma)^2}{9} \right]^2 \frac{|I_{t2}(\boldsymbol{\theta}_t)| \lambda_t^2}{\tau_t}, \left[3 - \frac{(1-2\gamma)^2}{9} \right] C_{t2} \max_{s \in \{t+1,\ldots,T\}} \left\{ \bar{c}_{t+1,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)| \lambda_s^2}{\tau_s} \right\} \right\} \\
&\quad = \left[3 - \frac{(1-2\gamma)^2}{9} \right]^2 \max_{s \in \{t,\ldots,T\}} \left\{ \bar{c}_{t,s} \frac{|I_{s2}(\boldsymbol{\theta}_s)| \lambda_s^2}{\tau_s} \right\}
\end{aligned}$$

where $\bar{c}_{t,t} = 1$ and

$$\begin{aligned}
\bar{c}_{t,s} &= 9 C_{t2} \bar{c}_{t+1,s} / [27 - (1-2\gamma)^2] \\
&= 2(T-t)^2 (S+1) \left[\frac{81 \max_{s \in \{t+1,\ldots,T\}} \{\bar{c}_{t+1,s} / c_{t+1,s}\}}{16(1-2\gamma)^2} + 1 \right] \left[3 - \frac{(1-2\gamma)^2}{9} \right] \bar{c}_{t+1,s},
\end{aligned}$$

for $s = t+1, \ldots, T$. In addition, since

$$\begin{aligned}
\max_{j,k \in \{1,\ldots,J_t\}} \left| (\mathbb{E}_n - E)\left(\frac{\phi_{tj} \phi_{tk}}{\bar{w}_{tj} \bar{w}_{tk}} \right) \right| &\le \frac{(1-2\gamma)^2 \max_{s \in \{t,\ldots,T\}} \{\bar{c}_{t,s} / c_{t,s}\} \lambda_t^2}{144 \max_{s \in \{t,\ldots,T\}} \left\{ \left[\max_{s \in \{t,\ldots,T\}} \{\bar{c}_{t,s} \lambda_t^2 / c_{t,s}\} \right] I_{s2}(\boldsymbol{\theta}_s) / \tau_s \right\}} \\
&\le \frac{(1-2\gamma)^2 \max_{s \in \{t,\ldots,T\}} \{\bar{c}_{t,s} / c_{t,s}\} \lambda_t^2}{144 \max_{s \in \{t,\ldots,T\}} \left\{ \bar{c}_{t,s} I_{s2}(\boldsymbol{\theta}_s) \lambda_s^2 / \tau_s \right\}},
\end{aligned}$$

it is easy to verify that

$$
\begin{aligned}
&\mathbb{E}_n[\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})]^2 \\
&\le \max_{j,k\in\{1,\ldots,J_t\}}\left|(\mathbb{E}_n-E)\Big(\frac{\phi_{tj}\phi_{tk}}{\bar{w}_{tj}\bar{w}_{tk}}\Big)\right|\Big(\sum_{j=J_{t1}+1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Big)^2+E[\Phi_{t2}^T(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})]^2 \\
&\le \left[\frac{81\max_{s\in\{t,\ldots,T\}}\{\bar{c}_{t,s}/c_{t,s}\}}{16(1-2\gamma)^2}+1\right]\left[3-\frac{(1-2\gamma)^2}{9}\right]^2\max_{s\in\{t,\ldots,T\}}\left\{\bar{c}_{t,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\right\}.
\end{aligned}
$$

This completes the proof. □

Proof of Lemma 6.

Note that $||\phi_{tj}\phi_{tk}/(\bar{w}_{tj}\bar{w}_{tk})-E[\phi_{tj}\phi_{tk}/(\bar{w}_{tj}\bar{w}_{tk})]||_\infty \le 2u^2$, and $E[\phi_{tj}\phi_{tk}/(\bar{w}_{tj}\bar{w}_{tk})]^2 \le b^2u^2$ for all $j,k\in\{1,\ldots,J_t\}$ by Assumption (B2). We apply Lemma S.1(a) with $\zeta_i = \pm[\phi_{tj}\phi_{tk}/(\bar{w}_{tj}\bar{w}_{tk})$
$-E(\phi_{tj}\phi_{tk}/(\bar{w}_{tj}\bar{w}_{tk})]$ and $s=(1-2\gamma)^2n/[144\max_{s\in\{t,\ldots,T\}}\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}]$. Using the union bound argument, we obtain

$$
\begin{aligned}
&\mathbf{P}(\{\Omega_{t,1}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\}^C) \\
&\le J_t(J_t+1)\exp\left(-\frac{(1-2\gamma)^4n}{2u^2\max_s\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}[144^2b^2\max_s\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}+96(1-2\gamma)^2]}\right) \\
&\le \exp(-\varphi)/3,
\end{aligned}
$$

where the second inequality follows from the definition of Θ_t in (B.2). □

Proof of Lemma 7.

We first show the result for $t=T$. For any $\boldsymbol{\theta}_T\in\Theta_T$, $\max_j\left|E\left[\Phi_T^\top(\boldsymbol{\theta}_T-\boldsymbol{\theta}_T^*)\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| \le \gamma\lambda_T b$ under Assumption (B4). Since $\boldsymbol{\theta}_T^*$ minimizes $E[Y_T-\Phi_T^\top\boldsymbol{\theta}_T]^2$, we have $E[(Y_T-\Phi_T^\top\boldsymbol{\theta}_T^*)\phi_{Tj}/\bar{w}_{Tj}] =$

0 for $j = \{1, \ldots, J_T\}$. Thus,

$$\max_{j\in\{1,\ldots,J_T\}} \left| E\left[\left(Y_T - \Phi_T^\top \boldsymbol{\theta}_T\right)\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| = \max_{j\in\{1,\ldots,J_T\}} \left| E\left[\Phi_T^\top(\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^*)\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| \le \gamma\lambda_T b.$$

This implies

$$\begin{aligned}\max_{j\in\{1,\ldots,J_T\}} &\left|\mathbb{E}_n\left[\left(Y_T - \Phi_T^\top \boldsymbol{\theta}_T\right)\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| \\ &\le \max_{j\in\{1,\ldots,J_T\}} \left|(\mathbb{E}_n - E)\left[\epsilon_T\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| + \max_{j\in\{1,\ldots,J_T\}} \left|(\mathbb{E}_n - E)\left[(Q_T^o - \Phi_T^\top\boldsymbol{\theta}_T)\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| + \gamma\lambda_T b.\end{aligned}$$

Under Assumptions (B1) and (B2), we have $E(\epsilon_{Ti}\phi_{Tj}/\bar{w}_{Tj}) = 0$ and $\sum_{i=1}^n E|(\epsilon_{Ti}\phi_{Tj}/\bar{w}_{Tj})^l| \le l!n\sigma^2 b^2(cu)^{l-2}/2$ for $j \in \{1, \ldots, J_T\}$ and all integers $l \ge 2$. Applying Lemma S.1(b), we obtain

$$\mathbf{P}\left(\left|(\mathbb{E}_n - E)\left[\epsilon_T\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| > \frac{1-2\gamma(3b-2)}{12}\lambda_T\right) \le 2\exp\left(-\frac{[1-2\gamma(3b-2)]^2\lambda_T^2 n}{288\sigma^2 b^2 + 24c[1-2\gamma(3b-2)]u\lambda_T}\right).$$

Similarly, by the definition of Θ_T and Assumption (B2), for any $\boldsymbol{\theta}_T \in \Theta_T$ and $j \in \{1, \ldots, J_T\}$, $\|(Q_T^o - \Phi_T^\top\boldsymbol{\theta}_T)\phi_{Tj}/\bar{w}_{Tj} - E((Q_T^o - \Phi_T^\top\boldsymbol{\theta}_T)\phi_{Tj}/\bar{w}_{Tj})\|_\infty \le 4\eta u$ and $E[(Q_T^o - \Phi_T^\top\boldsymbol{\theta}_T)\phi_{Tj}/\bar{w}_{Tj}]^2 \le 4\eta^2 b^2$. Then we have

$$\begin{aligned}\mathbf{P}\left(\left|(\mathbb{E}_n - E)\left[(Q_T^o - \Phi_T^\top\boldsymbol{\theta}_T)\frac{\phi_{Tj}}{\bar{w}_{Tj}}\right]\right| > \frac{1-2\gamma(3b-2)}{12}\lambda_T\right) \\ \le 2\exp\left(-\frac{[1-2\gamma(3b-2)]^2\lambda_T^2 n}{288(2\eta b)^2 + 32[1-2\gamma(3b-2)]u\eta\lambda_T}\right).\end{aligned}$$

The result follows from the union bound argument and condition (B.3).

Next, we show the results for $t < T$. For any $\boldsymbol{\theta}_s \in \Theta_s$, $s = t+1, \ldots, T$, note that

$$E\left[\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) - \tilde{Y}_t(\boldsymbol{\theta}_{t+1}^*, \ldots, \boldsymbol{\theta}_T^*)\right]^2$$

$$
\begin{aligned}
&= E\left\{ \sum_{s=t+1}^{T} \left[\max_{a_s} \Phi_s^\top(H_s, a_s)\boldsymbol{\theta}_s - \max_{a_s} \Phi_s^\top(H_s, a_s)\boldsymbol{\theta}_s^* - \Phi_s^\top(H_s, A_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right] \right\}^2 \\
&\le 2(T-t) \sum_{s=t+1}^{T} \left\{ E\left[\max_{a_s} \left[\Phi_s^\top(H_s, a_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right]^2 \right] + E\left[\Phi_s^\top(H_s, A_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right]^2 \right\} \\
&\le 2(T-t) \sum_{s=t+1}^{T} \left\{ E\left[\textstyle\sum_{a_s} \left[\Phi_s^\top(H_s, a_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right]^2 \right] + E\left[\Phi_s^\top(H_s, A_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right]^2 \right\} \\
&\le 2(T-t) \sum_{s=t+1}^{T} \left\{ E\left[\textstyle\sum_{a_s} p(a_s|H_s) S \left[\Phi_s^\top(H_s, a_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right]^2 \right] + E\left[\Phi_s^\top(H_s, A_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right]^2 \right\} \\
&\le 2(T-t) \sum_{s=t+1}^{T} \left\{ SE\left[\Phi_s^\top(H_s, A_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right]^2 + E\left[\Phi_s^\top(H_s, A_s)(\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^*) \right]^2 \right\} \\
&\le 2(T-t)(S+1)\gamma^2 \sum_{s=t+1}^{T} \lambda_s^2 \\
&\le \frac{9}{16}\gamma^2 \lambda_t^2,
\end{aligned}
$$

where the last inequality holds under the condition that $\lambda_t^2 \ge c_{t,s}\lambda_s^2$ and the fact that $c_{t,s} \ge 32(S+1)(T-t)^2/9$. Since $\boldsymbol{\theta}_t^*$ minimizes $E\left[\tilde{Y}_t(\boldsymbol{\theta}_{t+1}^*, \ldots, \boldsymbol{\theta}_T^*) - \Phi_t^\top \boldsymbol{\theta}_t\right]^2$, we have $E\left[\left(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}^*, \ldots, \boldsymbol{\theta}_T^*) - \Phi_t^\top \boldsymbol{\theta}_t^*\right)\phi_{tj}\right] = 0$. Thus, for $j = 1, \ldots, J_t$, we have

$$
\begin{aligned}
&\left| E\left[\left(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) - \Phi_t^\top \boldsymbol{\theta}_t \right) \frac{\phi_{tj}}{\bar{w}_{tj}} \right] \right| \\
&\quad \le \left| E\left[\left(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) - \tilde{Y}_t(\boldsymbol{\theta}_{t+1}^*, \ldots, \boldsymbol{\theta}_T^*) \right) \frac{\phi_{tj}}{\bar{w}_{tj}} \right] \right| + \left| E\left[\Phi_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*) \frac{\phi_{tj}}{\bar{w}_{tj}} \right] \right| \\
&\quad \le \frac{7}{4}\gamma b \lambda_t,
\end{aligned}
$$

where the last inequality holds from Assumption (B4). Hence,

$$
\begin{aligned}
&\max_{j \in \{1, \ldots, J_t\}} \left| \mathbb{E}_n \left[\left(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) - \Phi_t^\top \boldsymbol{\theta}_t \right) \frac{\phi_{tj}}{\bar{w}_{tj}} \right] \right| \\
&\qquad \le \max_{j \in \{1, \ldots, J_t\}} \left| (\mathbb{E}_n - E) \left[\left(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \ldots, \boldsymbol{\theta}_T) - \Phi_t^\top \boldsymbol{\theta}_t \right) \frac{\phi_{tj}}{\bar{w}_{tj}} \right] \right| + \frac{7}{4}\gamma b \lambda_t
\end{aligned}
$$

$$\leq \max_{j\in\{1,\ldots,J_t\}} \left|(\mathbb{E}_n - E)\left[\epsilon_t \frac{\phi_{tj}}{\bar{w}_{tj}}\right]\right| + \max_{j\in\{1,\ldots,J_t\}} \left|(\mathbb{E}_n - E)\left[f(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\frac{\phi_{tj}}{\bar{w}_{tj}}\right]\right| + \frac{7}{4}\gamma b\lambda_t,$$

where $f(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T) = Q_t^o(H_t,A_t) - \Phi_t^\top\boldsymbol{\theta}_t + \sum_{s=t+1}^T \left[-\max_{a_s} Q_s^o(H_s,a_s) + Q_s^o(H_s,A_s) - \Phi_s^\top\boldsymbol{\theta}_s + \max_{a_s}\Phi_s^\top\boldsymbol{\theta}_s\right]$.

Under Assumptions (B1) and (B2), we have $E(\epsilon_{ti}\phi_{tj}/\bar{w}_{tj}) = 0$ and $\sum_{i=1}^n E|(\epsilon_{ti}\phi_{tj}/\bar{w}_{tj})^l| \leq l!n\sigma^2b^2(cu)^{l-2}/2$ for all integers $l \geq 2$. Thus, for $\delta = (4\gamma+1)/12 - 7b\gamma/8$, if we apply Lemma S.1(b), we have

$$\mathbf{P}\left(\left|(\mathbb{E}_n - E)\left[\epsilon_{ti}\frac{\phi_{tj}}{\bar{w}_{tj}}\right]\right| > \delta\lambda_t\right) \leq 2\exp\left(-\frac{\delta^2\lambda_t^2 n}{2[\sigma^2b^2 + cu\delta\lambda_t]}\right).$$

Similarly, under Assumptions (B2) and (B4), $\left\|f(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\phi_{tj}/\bar{w}_{tj} - E(f(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\phi_{tj}/\bar{w}_{tj})\right\|_\infty \leq 4[1+2(T-t)]\eta u$ and $E[f(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\phi_{tj}/\bar{w}_{tj}]^2 \leq 4[1+2(T-t)]^2\eta^2b^2$. Applying Lemma S.1(a) yields

$$\mathbf{P}\left(\left|(\mathbb{E}_n - E)f(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\frac{\phi_{tj}}{\bar{w}_{tj}}\right| > \delta\lambda_t\right) \leq 2\exp\left(-\frac{\delta^2\lambda_t^2 n}{2\left[4[1+2(T-t)]^2\eta^2b^2 + 4[1+2(T-t)]\eta u\delta\lambda_t/3\right]}\right).$$

The result follows from the union bound argument and condition (B.4). □

Proof of Lemma 8.

Note that $||\sum_{a_t}\phi_{tj}(H_t,a_t)\phi_{tk}(H_t,a_t)/(\bar{w}_{tj}\bar{w}_{tk}) - E[\sum_{a_t}\phi_{tj}(H_t,a_t)\phi_{tk}(H_t,a_t)/(\bar{w}_{tj}\bar{w}_{tk})]||_\infty \leq 2|\mathscr{A}_t|u^2$, and $E[\sum_{a_t}\phi_{tj}(H_t,a_t)\phi_{tk}(H_t,a_t)/(\bar{w}_{tj}\bar{w}_{tk})]^2 \leq |\mathscr{A}_t|^2b^2u^2$ for all j,k at t-stage by Assumption (B2). Now we apply Lemma S.1(a) with

$$\zeta_i = \pm[\textstyle\sum_{a_{ti}}\phi_{tj}(H_{ti},a_{ti})\phi_{tk}(H_{ti},a_{ti})/(\bar{w}_{tj}\bar{w}_{tk}) - E(\sum_{a_{ti}}\phi_{tj}(H_{ti},a_{ti})\phi_{tk}(H_{ti},a_{ti})/(\bar{w}_{tj}\bar{w}_{tk})]$$

and $s = (1-2\gamma)^2|\mathscr{A}_t|n/[144\max_{s\in\{t,\ldots,T\}}\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}]$,

where $\zeta_i \le q = 2|\mathscr{A}_t|u^2$ and $\sum_{i=1}^n E\zeta_i^2 \le \nu = n|\mathscr{A}_t|^2b^2u^2$. Using the union bound argument, we obtain

$$\begin{aligned}
&\mathbf{P}(\{\Omega_{t,3}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_T)\}^C)\\
&\le J_t(J_t+1)\exp\left(-\frac{(1-2\gamma)^4 n}{2u^2\max_s\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}[144^2b^2\max_s\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}+96(1-2\gamma)^2]}\right)\\
&\le \exp(-\varphi)/3,
\end{aligned}$$

where the second inequality follows from the definition of Θ_t in (B.2). □

Lemma S.1. ***(Bernstein's inequalities)*** *Let $\zeta_1,\ldots,\zeta_n$ be independent and square integrable random variables such that $E(\zeta_i)=0$ for all $i=1,\ldots,n$.*

(a) Assume there exists some positive constants q and ν such that $\zeta_i \le q$ a.s. for all $i=1,\ldots,n$ and $\sum_{i=1}^n E\zeta_i^2 \le \nu$. Then for any $s>0$,

$$\mathbf{P}\Big(\sum_{i=1}^n \zeta_i > s\Big) \le \exp\Big(-\frac{s^2}{2(\nu+qs/3)}\Big).$$

(b) Assume there exists some positive constants q and ν such that $\sum_{i=1}^n E[(\zeta_i^l)_+] \le l!\nu q^{l-2}/2$ for all $l\ge 2$. Then for any $s>0$,

$$\mathbf{P}\Big(\sum_{i=1}^n \zeta_i > s\Big) \le \exp\Big(-\frac{s^2}{2(\nu+qs)}\Big).$$