Statistical Inference for Adaptive Experimenta

HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences
have examined a dissertation entitled:

"Statistical Inference for Adaptive Experimentation"

presented by: Kelly W. Zhang

Signature _____
Typed name: Professor S. Murphy

Signature _____
Typed name: Professor L. Janson

Signature _____
Typed name: Professor M. Tambe

Signature _____
Typed name: Professor J. O. Jonasson

April 26, 2023

Statistical Inference for Adaptive Experimentation

A DISSERTATION PRESENTED
BY
KELLY W. ZHANG
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2023

Thesis advisors: Susan A. Murphy and Lucas Janson Kelly W. Zhang

Statistical Inference for Adaptive Experimentation

Abstract

Online reinforcement learning (RL) algorithms are a very promising tool for personalizing decision-making for digital interventions, e.g., in mobile health, online education, and public policy. Online RL algorithms are increasingly being used in these applications since they are able to use previously collected data to continually learn and improve future decision-making. After deploying online RL algorithms though, it is critical to be able to answer scientific questions like: Did one type of teaching strategy lead to better student outcomes? In which contexts is a digital health intervention effective? The answers to these questions inform decisions about whether to roll out or how to improve a given intervention. Constructing confidence intervals for treatment effects using normal approximations is a natural approach to address these questions. However, classical statistical inference approaches for i.i.d. data fail to provide valid confidence intervals on data collected with online RL algorithms. Since online RL algorithms use previously collected data to inform future treatment decisions, they induce dependence in the collected data. This induced dependence can cause standard statistical inference approaches for i.i.d. data to be invalid on this data type. This thesis provides an understanding of the reasons behind the failure of classical methods in these settings. Moreover, we introduce a variety of alternative statistical inference approaches that are applicable to data collected by online RL algorithms.

Contents

List of Figures

Acknowledgments

First and foremost I'd like to thank my advisors Susan Murphy and Lucas Janson. I feel extremely lucky and grateful to have them both as my advisors. Working with them has deeply impacted my growth and trajectory as a researcher, and a person in general. From this experience working with Susan and Lucas, I am inspired in my career to work on choosing impactful research questions, especially ones that arise from working on interdisciplinary collaborative projects.

I want to thank the following friends and mentors who helped me first get into research, especially Sasha Rush, Sam Bowman, Yoon Kim, Adji Dieng, Jake Zhao, Tian Wang, Will Whitney, and Mikael Henaff.

I would like to give special thanks to Preetum Nakkiran for interesting research discussions and for introducing me to the TCS community at Harvard; Raaz Dwivedi for his friendship, advice, and support during the last two years of my PhD; Anna Trella for being a great collaborator and for our fun times working together; and Jayshree Sarathy for our three years living together and going through the PhD journey with me.

More broadly, I want to thank the great mentors and friends I've had the chance to work with and learn from, including, Mash Rabbi, Walter Dempsey, Tianchen Qian, Peng Liao, Shuangning Li, Kyra Gan, Yongyi Guo, Jens Winkowski, Moritz Granule, Ben Edelman, Thibault Horel, Lu Zhang, Omer Gottesman, Jiayu Yao, Prayaag Venkat, Boriana Gjura, Finale Doshi-Velez, Billie Inbal-Shani, Vivek Shetty, Sean Jewell, Nick Foti, and Guillermo Sapiro.

I also want to thank the following people for advice and guidance during my last year when I was looking for a job: Daniel Russo, Hongseok Namkoong, Nathan Kallus, Stefan Wager, Aaditya Ramdas, Sonali Parbhoo, Ciara Pike-Burke, and Joseph Jay Williams.

Finally, I thank my family for their support, particularly my dad, Michael, and Frank.

1

Introduction

Online reinforcement learning (RL) algorithms are a key tool for optimizing and improving decision-making over time. As our lives become increasingly digitized, there are more opportunities to use online RL algorithms to optimize decision-making in socially beneficial endeavors—including digital health, online education, and public policy. For example, in mobile health, online RL algorithms are used to optimize the delivery of personalized messages through smartphones and wearables to help users move closer towards their health goals[67,105,96,97,34,38]. In public policy settings, text mes-

sage interventions using online RL algorithms are being used to help optimize reminders to individuals awaiting trial to attend their court hearings[91]. As a result, there are more studies in these social science domains using online RL algorithms. Theory for online RL algorithms provides guarantees as to how well the algorithms are able to optimize decision-making during the course of the study.

For these types of real-world studies deploying online RL algorithms, it is critical to be able to evaluate the effectiveness of these deployments after the study concludes. Specifically, it is important to be able to answer scientific questions like: Did one type of teaching strategy lead to better student outcomes? In which contexts is a digital health intervention effective? The answers to these questions inform decisions about whether to roll out or how to improve a given intervention. Additionally, this information can be used to publish and share results, as well as inform stakeholders.

The typical approach for addressing these types of scientific questions is to construct constructing a confidence interval for treatment effects, e.g., the expected difference in reward under two different treatments in a particular state. However, standard approaches for constructing these confidence intervals do not accurately capture the confidence when applied to data collected by online RL and other adaptive algorithms. Online RL algorithms learn to improve decision making by using previously collected data from users to inform future decision making on other users. This leads the data collected across users to be dependent through the algorithm. This dependence violates the independence assumptions that underlie many standard statistical inference approaches. In fact, this thesis shows that on data collected by common online RL algorithms, standard approaches for constructing confidence intervals for treatment effects can provably, consistently be overconfident. [109]

Social scientists are often unwilling and scared of using online RL algorithms in practice if they cannot ensure that they can construct valid confidence for intervals for treatment effects after the study is over. Developing reliable, general methods for statistical inference on adaptively collected data is critical to ensure that online RL algorithms can be deployed without sacrificing the reliability of the evaluation of treatment effects. This is especially important now with current climate of

mistrust in science to practice responsible data science.

This thesis first develops a theoretical understanding of the reasons behind the failure of standard methods to construct confidence intervals on data collected with RL algorithms. We then proceed to use this understanding to develop a variety of methods for constructing confidence intervals on data collected by RL algorithms in a variety of environments—including bandit, contextual bandit, and longitudinal (non-Markovian) environments. This work has a significant impact on our theoretical understanding of the statistical inference for data collected by online RL algorithms. Moreover, this work also has already had a significant impact on real-world adaptive experiments utilizing online RL algorithms. We provide an overview of these real-world deployments in the next section.

1.1 Impact (Use Cases)

In just the past few years, the statistical methods developed in this thesis have already impacted how researchers (beyond just direct collaborators of mine) analyze the data collected by experiments using online RL and other adaptive algorithms. We provide an overview of three different deployments of online RL that have used (or plan to use) statistical methods developed in this thesis for the after study data analysis. The first example discusses a direct collaborations of mine, and the following two describe adaptive experiments run by others that used our statistical methods.

1.1.1 Digital (Mobile) Health

The treatment for many chronic health problems like hypertension, addition, and mental illness, involve patients changing their behavior. For example, these behaviors could include exercising, or practicing mindfulness or a cognitive behavioral therapy skill. Digital and mobile health interventions are a way to help users develop healthier habits in an "in-the-moment" fashion, that typically a

doctor cannot personally provide in between visits.

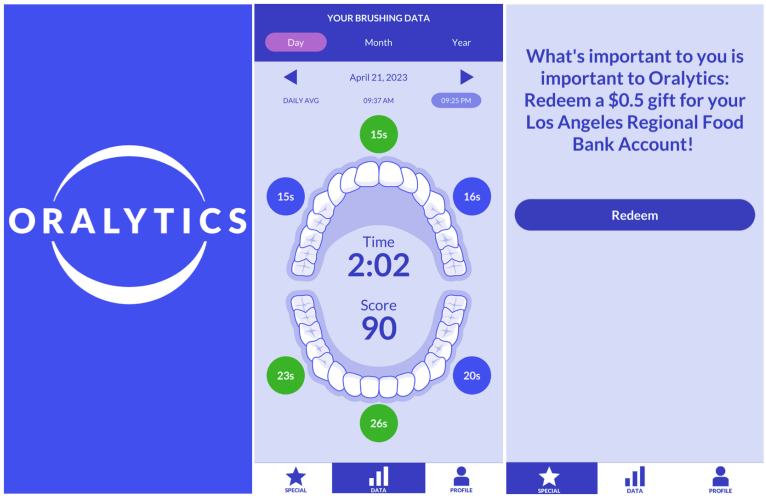

Figure 1.1: Panels from the Oralytics app. The second panel displays the user's past brushing data. The third panel shows an example engagement push message from the app; the message is designed to take advantage of the behavioral science concept of "reciprocity" (see Nahum-Shani et al., 2023[74]).

For the past several years, I've been a part of Oralytics,[96,97] an interdisciplinary collaboration developing a mobile health intervention to help users develop better toothbrushing habits. Oral diseases are largely preventable through regular brushing and flossing.* Despite this, according to the World Health Organization 5-10% of healthcare budgets in industrialized countries are spent on treating dental cavities,[1] and according to the US CDC nearly one-fifth of U.S. adults 65 or older have lost all their teeth.[30] The Oralytics study participants will have a blue-tooth enabled toothbrush and a mobile app. The app is designed to augment standard dental care between dentist visits, by sending users push messages at opportune moments to encourage brushing.

The Oralytics app will use a reinforcement learning algorithm to optimize the delivery of these intervention messages. Specifically, the RL algorithm will use user state information collected from the blue-tooth enabled toothbrush and data collected from the app, decide whether or not to send

*World Health Organization: https://www.who.int/news-room/fact-sheets/detail/oral-health

the message at that moment. Each participant is in the study for 10 weeks and there are two decision times a day, one before the morning and evening brush times respectively. The algorithm is designed to optimize decision making to maximize the users' brushing quality over time, as measured by the blue-tooth toothbrush.

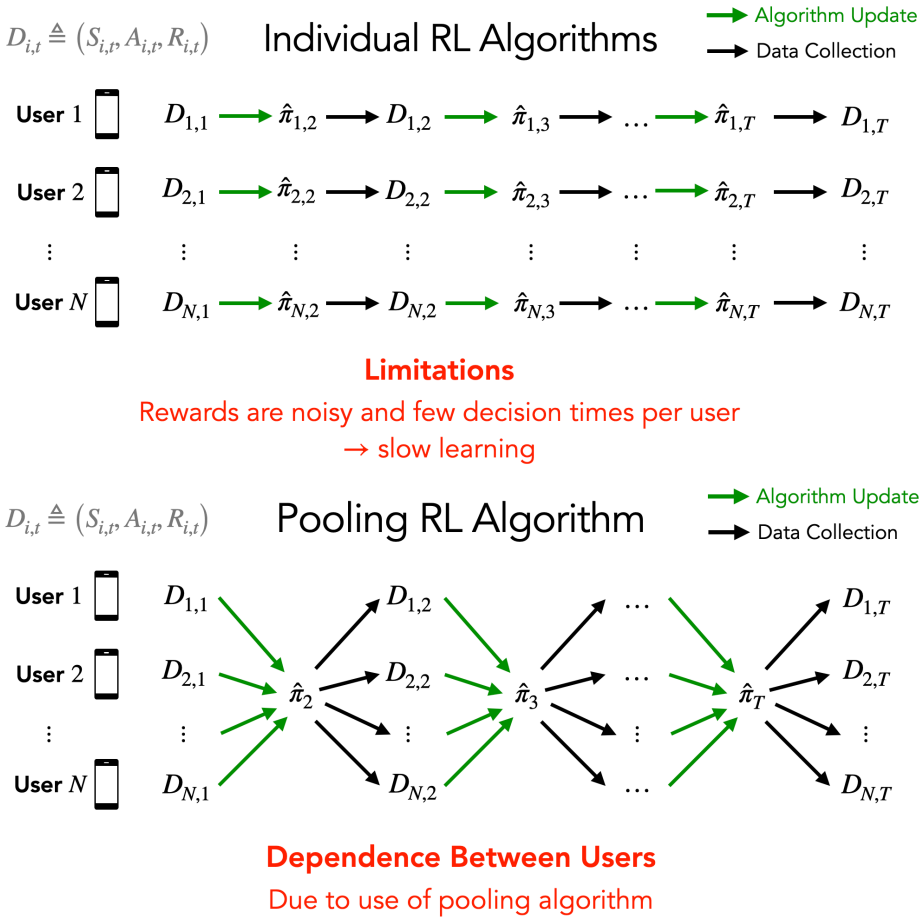

Figure 1.2: Illustrating the design of studies using individual RL algorithms (above) versus pooling RL algorithms (below).

Since the RL algorithm will have relatively few data points to learn from (140 decision times = 10 weeks $\times$7 days a week $\times$2 decision times a day), we really wanted to use an RL algorithm

that was able to learn using the data of multiple users to learn. In simulation environments built using previously collected brushing data, we found that using an RL algorithm that combined or "pooled" data across users to learn led to significantly higher rewards than using a separate RL algorithm per user; see Trella et al., Algorithms 2022.[96] If the RL algorithm is able to use the data of multiple users to inform decision making, this leads the data trajectories collected from the users to be dependent through the algorithm; see Figure 1.2 for illustration. However, there we no existing methods for constructing confidence intervals for treatment effects on data collected by such algorithms which cause the user data trajectories to be dependent. The Oralytics trial is a micro-randomized trial[60,80] that is registered with clinicaltrials.gov[†]. Ensuring we can construct confidence intervals for treatment effects is critical to ensure replicable and responsible science.

Ultimately, the work Zhang, Janson, & Murphy, 2023 [111] one of the works presented in this thesis was able to address this challenge. Specifically, this work developed the Adaptive Sandwich Variance estimator, an adjustment to the standard Sandwich Variance estimator, that accounts for the adaptive sampling. This work *enabled* my collaborators and I to run a pooling RL algorithm for the Oralytics study. The study is currently in the pilot phase and the micro-randomized trial phase will start in summer 2023.

1.1.2 Public Policy

In public policy and economic experiments, adaptive algorithms are of interest to better optimize treatment allocations. The focus is often on using these algorithms to learn a best policy after the study concludes. For example, the following work describes an adaptive policy experiment that was run by the World Bank:

Bruno Esposito and Anja Sautmann. "Adaptive Experiments for Policy Choice: Phone Calls for

[†]Micro-Randomized Trial to Optimize Digital Oral Health Behavior Change Interventions: https://clinicaltrials.gov/ct2/show/NCT05624489?term=oralytics&draw=2&rank=1

Home Reading in Kenya," *World Bank Group.*

The goal of the experiment was to understand how phone calls with Interactive Voice Response (IVR) technology could be used to encourage parents in Kenya to read with their children. The experiment was run in partnership with NewGlobe, an organization that works with governments to develop technology-enabled education systems. The experiment involved using an Bayesian exploration sampling adaptive algorithm to choose between six different IVR call types. Each IVR call type provided parents with reading exercises to practice with their children, but the delivery approach and difficulty of the exercises varied.

The main outcome the algorithm targeted was parental engagement with the call; they also assess the effect of parental engagement on the downstream outcome of interest, children's reading fluency. This experiment used the adaptively weighted M-estimator approach from Zhang, Janson, & Murphy, NeurIPS 2021 [110], one of the works presented in this thesis, for constructing confidence intervals for the expected outcomes under each treatment. They found that IVR calls that gave parents reading exercises to practice with their child later, were more engaging than those that asked the parent to join the call with the child to practice. Additionally, they found that IVR calls that either gave intermediate level exercises or allowed parents to choose the difficulty of exercises performed significantly better than IVR calls that chose the difficulty based on the child's baseline measures.

1.1.3 Political Science

The following work describes the use of adaptive algorithms in political science experiments:

Molly Offer-Westort, Alexander Coppock, Donald P. Green. "Adaptive Experimental Design: Prospects and Applications in Political Science," *American Journal of Political Science.*

They discuss how political science experiments often have many different treatments and adaptive algorithms can be used to assign larger treatment probabilities to the most promising treatments

to better discover the best performing treatment.[77] They find in simulations that using adaptive algorithms, rather than uniform sampling can significantly improve the probability that they can find a significantly high performing treatment.

In their work, they actually run an adaptive experiment to understand different approaches to reducing peoples' political bias when answering factual questions. Financial incentives, urging respondents to put aside political biases, and allowing respondents more time to reflect and consult outside sources are all plausible ways to reduce their political bias and increase their accuracy in responding to factual questions. They run an adaptive trial using a posterior sampling algorithm with six different treatments. The reward is the accuracy of the respondents' answers to factual questions on budget deficits, black unemployment, and differences in farm industry income in the years under Trump vs Obama. At each decision time, a batch of about 300 respondents were a treatment; there were a total of 10 batches, so about 3000 survey takers total.

To analyze the results of their experiment, they construct confidence intervals for treatment effects for the best performing treatment compared to a control condition. Specifically, they use the batched OLS approach from Zhang, Janson, & Murphy, NeurIPS 2020[109], one of the works presented in this thesis, for constructing confidence intervals. In the end, they found the "lottery" condition in which respondents are given an extra chance to win an Amazon gift card if they answer correctly led both democrats and republicans to answer the most accurately. Their work provides an early example as to how adaptive algorithms can be used to in political science survey experiments to increase the probability of finding a significantly high-performing treatment.

1.2 Thesis Outline

The remainder of this thesis is organized as follows:

- Chapter 2 first provides a brief overview of reinforcement learning and standard methods for

constructing confidence intervals. We then provide an overview of the challenges of statistical inference on data collected with online RL algorithms.

- Chapter 3 develops theory for constructing confidence intervals on data collected in multi-arm and contextual bandit environments in which RL algorithms select treatments in batches. We prove that the ordinary least squares estimator (OLS), which is asymptotically normal on independently sampled data, is not asymptotically normal on data collected using common bandit algorithms. Thus, the naive assumption that the OLS estimator is approximately normal can lead to consistently overconfident confidence intervals. We introduce the Batched OLS estimator that we prove is asymptotically normal on (contextual) bandit data, and is robust to non-stationarity in the baseline reward. This chapter is based on work from the following publication:

  Kelly W. Zhang, Lucas Janson, Susan A Murphy. "Inference for Batched Bandits," *NeurIPS 2020*.

- Chapter 4 develops a general method for conducting statistical inference using more complex models on data collected with (contextual) bandit algorithms. For example, previous methods cannot be used for valid inference on parameters in a logistic regression model for a binary reward. We develop theory justifying the use of M-estimators on (contextual) bandit data. Specifically, we show that M-estimators, modified with particular adaptive weights, can be used to construct asymptotically valid confidence regions for a variety of inferential targets. This chapter is based on work from the following publication:

  Kelly W. Zhang, Lucas Janson, Susan A Murphy. "Statistical Inference with M-Estimators on Adaptively Collected Data," *NeurIPS 2021*.

- Chapter 5 develops methods for constructing confidence intervals for treatment effects for longitudinal data collected from multiple users by RL algorithms that learn online by combining or "pooling" data across multiple users since this can lead potentially to faster learning. However, by pooling, these algorithms induce dependence between the collected user data trajectories; this makes statistical inference on this data type especially challenging. We provide methods to perform a variety of statistical analyses on such adaptively collected data, including Z-estimation and inferring excursion effects. For these results, we develop novel empirical process theory for non-i.i.d., adaptively collected data. This chapter is based on work from the following publication:

  Kelly W. Zhang, Lucas Janson, Susan A Murphy. "Statistical Inference After Adaptive Sampling for Longitudinal Data," *Under submission.*

- Finally, chapter 6 concludes and discusses open questions.

2

Background

2.1 Reinforcement Learning

We now describe the sequential decision-making problems that we use RL algorithms for. Let there be T decision times total. We use $t \in [1 : T]$ to index these decision times.

State At each decision time t, we observe some state or context information that can be used to tailor the decision, which we denote using $S_t \in \mathcal{S}$. For example, in mobile health, this could be

sensor data collected from the user's wearable or mobile phone, and in online education, this could be the user's previous performance on homework, as well as demographic or browser information from the user.

ACTION The action (or treatment) A_t is assumed to be in some action space $\mathcal{A}$. For example, in a binary action setting, $\mathcal{A} = \{0, 1\}$. In mobile health, the action could correspond to sending versus not sending the user a message, and in online education, the action could be different types of teaching strategies.

REWARD The reward $R_t \in \mathbb{R}$ is some outcome observed after the action A_t is taken. This is often some outcome one wants to maximize. For example, in mobile health, this could be the physical activity of the user following a decision time, and in online education, this could be student performance on homework or quizzes.

REINFORCEMENT LEARNING ALGORITHMS At a given decision time t, the RL algorithm is able to use the history of data observed so far, $\mathcal{H}_{t-1} := \left\{S_{t'}, A_{t'}, R_{t'}\right\}_{t'=1}^{t-1}$, to inform its future action selection. In particular, these reinforcement learning algorithms use $\mathcal{H}_{t-1}$ to form a **policy** π_t, which is a mapping from the state space $\mathcal{S}$ to a distribution over the actions space $\mathcal{A}$. RL algorithms are commonly designed to update and form policies such that they bias towards selecting actions that will lead to higher future rewards. In particular, these algorithms are designed to maximize the expected sum of rewards:

$$\mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T} R_t\right].$$

Above the expectation is indexed by $\mathcal{A}$ to denote that the actions are selected according to the policies formed by the reinforcement learning algorithm $\mathcal{A}$.

Additionally, in other problem settings, RL algorithms may be designed to maximize not the

expected sum of rewards collected during the experimental period; rather they are designed to maximize expected reward of a policy learned from the experimental data when deployed more widely. For example, the data collected during the study, $\mathcal{H}_T$, is used to form a policy $\hat{\pi}$. The goal of the algorithm is to collect data to maximize the expected reward $\hat{\pi}$ would achieve:

$$\mathbb{E}_{\hat{\pi}}\left[\sum_{t=1}^{T} R'_t\right].$$

Above use R'_t to denote the rewards the policy would receive if deployed on a new sample from the same environment the R_t rewards are generated from. Note that the above quantity is a random quantity, due to the randomness in the learned policy $\hat{\pi}$. Commonly one aims to maximize the above quantity with high probability; this objective is closely related to the best-arm identification literature.

Sequential Decision-Making Environments We use potential outcomes to represent counterfactual outcomes. The simplest sequential decision-making environment is that of a "bandit" in which the following outcomes are independent and identically distributed (i.i.d.) over decision times $t \in [1\colon T]$:

$$\left(S_t, \left\{R_t(a) : a \in \mathcal{A}\right\}\right). \tag{2.1.1}$$

More complicated sequential decision-making environments (Markovian and non-Markovian environments) are more complicated in that they (1) can allow actions to affect future state and reward distributions, and (2) there may be non-stationarity in the state and reward distributions over time.

2.2 Standard Approaches to Uncertainty Quantification

In this work, we focus on estimating and constructing confidence intervals for quantities like treatment effects. We start by discussing a simple example and then discuss more general cases.

2.2.1 Simple Difference in Means Examples

In "bandit" environments from display (2.1.1), a simple treatment effect of interest could be the difference in expected rewards under two different actions (for simplicity we consider $\mathcal{A} = \{0, 1\}$):

$$\Delta^\star \triangleq \mathbb{E}[R_t(1) - R_t(0)].$$

A simple estimator of the above treatment effect is the difference in sample means under each action:

$$\hat{\Delta} \triangleq \frac{\sum_{t=1}^T R_t A_t}{\sum_{t=1}^T A_t} - \frac{\sum_{t=1}^T R_t(1 - A_t)}{\sum_{t=1}^T (1 - A_t)}. \tag{2.2.1}$$

A common approach to constructing confidence intervals is to use the asymptotic distribution of estimators to approximate their finite sample distribution (this has a long history of success in science). For example, assuming $\sigma^2 = \mathrm{Var}\big(R_t(0)\big) = \mathrm{Var}\big(R_t(1)\big)$, if actions A_t are independently drawn from Bernoulli(0.5), then $\hat{\Delta}$ is asymptotically normal:

$$\sqrt{\frac{\big[\sum_{t=1}^T A_t\big]\big[\sum_{t=1}^T (1 - A_t)\big]}{T}} \left(\hat{\Delta} - \Delta^\star\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

We can use the above convergence result to construct an approximate, asymptotically valid 95% two-sided confidence interval for $\Delta^\star$:

$$\left[\hat{\Delta} - z_{0.975}\sigma\sqrt{\frac{T}{\big[\sum_{t=1}^T A_t\big]\big[\sum_{t=1}^T (1 - A_t)\big]}},\ \hat{\Delta} + z_{0.975}\sigma\sqrt{\frac{T}{\big[\sum_{t=1}^T A_t\big]\big[\sum_{t=1}^T (1 - A_t)\big]}}\right].$$

Above $z_{0.975}$ is the 0.975th quantile of the standard normal distribution.

2.2.2 Inference for More General Models

More generally, we may be interested in the parameters of a model of an outcome (which could be the reward). For example, here are some example outcome models for a binary action $\mathcal{A} = \{0, 1\}$ setting:

- Linear Model for Continuous Outcomes:

$$\mathbb{E}\left[R_t | S_t, A_t\right] = S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star$$

- Logistic Model for Binary Outcomes:

$$\mathbb{E}\left[R_t | S_t, A_t\right] = \left[1 + \exp\left(S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star\right)\right]^{-1}$$

- Poisson Model for Count Outcomes:

$$\mathbb{E}\left[R_t | S_t, A_t\right] = \log\left[S_t^\top \theta_0^\star + A_t S_t^\top \theta_1^\star\right]$$

In the models above, the parameter $\theta_1^\star$ parameterizes the treatment effect portion of the model, i.e., the part that parameterizes $\mathbb{E}\left[R_t | S_t, A_t = 1\right] - \mathbb{E}\left[R_t | S_t, A_t = 0\right]$.

A common approach to forming estimators $\hat{\theta}$ of these model parameters is use the minimizer of an empirical loss function:

$$\hat{\theta} \triangleq \operatorname{argmin} \frac{1}{T} \sum_{t=1}^{T} \ell_\theta\big(R_t,\ S_t,\ A_t\big).$$

For example, sample means, least squares, logistic regression, and maximum likelihood estimators can all be written as minimizers of empirical loss functions for different choices of ℓ.

Under regularity conditions (see Theorems 5.21 and 5.23 of Van der Vaart, 1996[100]), the minimizers of these loss functions converge in distribution to a normal with a sandwich variance:

$$\sqrt{n}\left(\hat{\theta}-\theta^{\star}\right) \xrightarrow{D} \\ \mathcal{N}\left(0, \mathbb{E}\left[\ddot{\ell}_{\theta^{\star}}\left(R_t, S_t, A_t\right)\right]^{-1} \mathbb{E}\left[\dot{\ell}_{\theta^{\star}}\left(R_t, S_t, A_t\right) \dot{\ell}_{\theta^{\star}}\left(R_t, S_t, A_t\right)^{\top}\right] \mathbb{E}\left[\ddot{\ell}_{\theta^{\star}}\left(R_t, S_t, A_t\right)\right]^{-1,\top}\right).$$

Above we use $\dot{\ell}$ and $\ddot{\ell}$ to refer to the first and second derivatives of the function ℓ with respect to θ, i.e., $\dot{\ell}_{\theta^{\star}}\left(R_t,\, S_t,\, A_t\right) \triangleq \frac{\partial}{\partial \theta} \ell_{\theta}\left(R_t,\, S_t,\, A_t\right)\big|_{\theta=\theta^{\star}}$ and $\ddot{\ell}_{\theta^{\star}}\left(R_t,\, S_t,\, A_t\right) \triangleq \frac{\partial^2}{\partial \theta \partial \theta^{\top}} \ell_{\theta}\left(R_t,\, S_t,\, A_t\right)\big|_{\theta=\theta^{\star}}$. The above convergence result can be used to construct approximate, asymptotically valid confidence regions for the treatment effect parameters $\theta_1^{\star}$.

2.3 Challenges for Statistical Inference on Adaptively Collected Data

2.3.1 Warm-Up: Simple Example Showing How Sample Means can be Biased in Small Samples

"Adaptively collected" data is data that is collected such that the decision about what to sample next depends on the results of previous samples. For example, in a multi-arm bandit setting, data collected by strategies like ε-greedy and Thompson sampling are adaptive, since these strategies sample treatments that gave high rewards in the past more than treatments that previously gave lower rewards. Adaptively collected data have a temporal dependency and thus the samples are not independent. The dependency between samples makes constructing unbiased estimators more difficult for adaptively collected data than for i.i.d data.

We describe one illustrative example about how the dependency between adaptively collected samples can lead to biased estimates when using estimators that assume i.i.d. data. Specifically, we will show how the sample mean of adaptive data can biased[76]. Note though that this bias is a finite

sample phenomena and will go to zero as the sample size increases.

Suppose we have two identical treatments that emit rewards that are drawn i.i.d. from a standard normal distribution. We draw a total of three rewards from these treatments: R_1, R_2, R_3. R_1 is always sampled from treatment 1 and R_2 is always sampled from treatment 2. We follow a greedy strategy and sample R_3 from the treatment that emits a higher reward from the first two samples. Let M_1 and M_2 represent the sample means of treatments 1 and 2 respectively.

If $R_1 > R_2$:

$$\begin{cases} M_1 = \frac{1}{2}(R_1 + R_3) \\ M_2 = R_2 \end{cases}$$

else $(R_1 > R_2)$:

$$\begin{cases} M_1 = R_1 \\ M_2 = \frac{1}{2}(R_2 + R_3) \end{cases}$$

We let $Z_1, Z_2, Z_3 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and use the following order statistics notation $Z_{(1)} \sim \min(Z_1, Z_2)$ and $Z_{(2)} \sim \max(Z_1, Z_2)$.

$$M_1 \sim \mathbb{P}(R_1 > R_2)\frac{Z_{(2)} + Z_3}{2} + \mathbb{P}(R_1 < R_2)Z_{(1)}$$

Thus, since by symmetry $\mathbb{P}(R_1 > R_2) = \mathbb{P}(R_1 < R_2) = \frac{1}{2}$,

$$\mathbb{E}[M_1] = \mathbb{P}(R_1 > R_2)\mathbb{E}[M_1 | R_1 > R_2] + \mathbb{P}(R_1 < R_2)\mathbb{E}[M_1 | R_1 < R_2]$$

$$= \mathbb{P}(R_1 > R_2)\mathbb{E}\left[\frac{Z_{(2)} + Z_3}{2}\right] + \mathbb{P}(R_1 < R_2)\mathbb{E}[Z_{(1)}]$$

$$= \frac{1}{2}\mathbb{E}\left[\frac{Z_{(2)} + Z_3}{2}\right] + \frac{1}{2}\mathbb{E}[Z_{(1)}]$$

$$= \frac{1}{4}\mathbb{E}[Z_{(2)}] + \frac{1}{2}\mathbb{E}[Z_{(1)}] < 0$$

since by symmetry, $\mathbb{E}[Z_{(2)}] = -\mathbb{E}[Z_{(1)}]$.

Thus we have shown in this simple three sample case, that the sample mean is negatively biased when using a greedy strategy. Note that by symmetry, if we use an "anti-greedy" strategy of sampling, in which we sample the treatment that gives a *lower* reward more, the expected sample mean *overestimates* the true mean.

Nie et al., AISTATS 2018[76] show that for bandit algorithms that have the properties "exploit" and "independence of irrelevant options", the sample mean will be negatively biased. The exploit property means that if treatment k is chosen, then in an alternative sample history, if the sample mean of treatment k were higher, then treatment k would still be chosen. The independence of irrelevant options property means that if treatment k is not chosen, which other treatment is selected only depends on the histories of the other treatments.

2.3.2 Uncertainty Quantification Challenges

The dependence induced by adaptively collected data also causes challenges for constructing confidence intervals. We consider the setting as described earlier in Section 2.2.1 and look the difference in sample means.

In Figure 2.1, we plot the empirical distribution of Z-statistics for $\hat{\Delta}$, the difference in sample means as defined in display (2.2.1). In the left hand plot data is collected independently with $A_t \sim$ Bernoulli(0.5) and on the right hand plot data is collected with a Thompson Sampling, a common Bayesian RL algorithm. From the figure, we see that the Z-statistic under Thompson Sampling is no longer asymptotically normal. This means that using a normal distribution to approximate the distribution of $\hat{\Delta}$ under Thompson Sampling will lead to unreliable confidence intervals that provide insufficient coverage. Thus, we cannot simply use standard statistical inference approaches

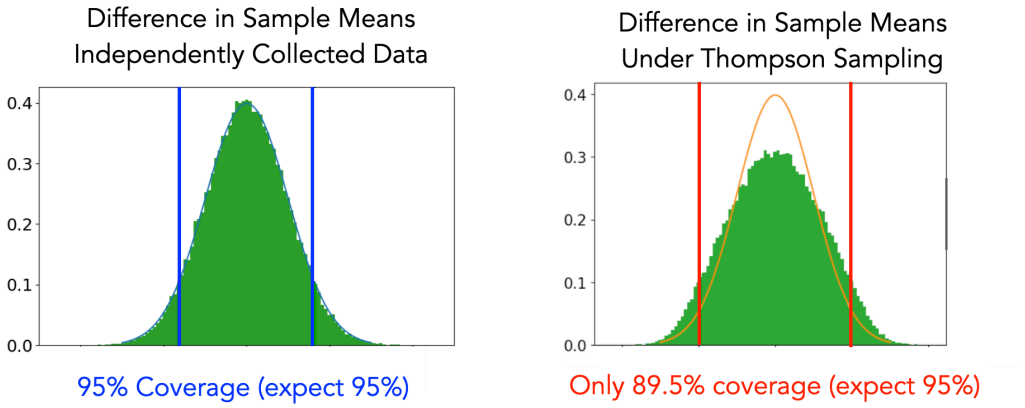

Figure 2.1: Z-statistic for difference in sample means in two-arm bandit setting with $T = 1000, 0 = \mathbb{E}[R_t(0)] = \mathbb{E}[R_t(1)]$, and $\mathcal{N}(0, 1)$ reward errors. Left side figure is data collected independently with $A_t \sim \text{Bernoulli}(0.5)$. Right side figure is data collected with Thompson Sampling with standard normal priors on each treatment.

on data collected with RL algorithms.

2.3.3 Desiderata

We now discuss what are the properties we want from a statistical inference approach for adaptively collected data.

- **Inference for Parameters of General Models**: First we would like to be able to infer the parameters in a variety of models (see examples discussed in Section 2.2.2).
- **Accommodates Variety of Common RL Algorithms**: We would like our statistical inference approach to be applicable to data collected by a variety of common RL algorithms, e.g., posterior sampling, ε-greedy, and UCB.
- **Applicable to Relevant Real-World Problem Settings**: Lastly, we would like our statistical inference approaches to be applicable to the types of real-world problem settings described in Section 1.1. These environments included those that correspond to classical multi-arm and contextual bandits, as well as longitudinal data settings.

These desiderata listed above guided the development of the statistical inference methods we describe in the following chapters of this thesis.

3

Inference for Batched Bandits

As bandit algorithms are increasingly utilized in scientific studies and industrial applications, there is an associated increasing need for reliable inference methods based on the resulting adaptively-collected data. In this work, we develop methods for inference on data collected in batches using a bandit algorithm. We first prove that the ordinary least squares estimator (OLS), which is asymptotically normal on independently sampled data, is *not* asymptotically normal on

data collected using standard bandit algorithms when there is no unique optimal arm. This asymptotic non-normality result implies that the naive assumption that the OLS estimator is approximately normal can lead to Type-1 error inflation and confidence intervals with below-nominal coverage probabilities. Second, we introduce the Batched OLS estimator (BOLS) that we prove is (1) asymptotically normal on data collected from both multi-arm and contextual bandits and (2) robust to non-stationarity in the baseline reward.

3.1 Introduction

Due to their regret minimizing guarantees bandit algorithms have been increasingly used in in real-world sequential decision-making problems, like online advertising[66], mobile health [105], and online education [82]. However, for many real-world problems it is not enough to just minimize regret on a particular problem instance. For example, suppose we have run an online education experiment using a bandit algorithm where we test different types of teaching strategies. When designing a new online course, ideally we could use the data from the previous experiment to inform the design, e.g., under-performing arms could be eliminated or modified. Moreover, to help others designing online courses we would like to be able to publish our findings about how different teaching strategies compare in their performance. This example donstrates the need for statistical inference methods on bandit data, which allow practitioners to draw generalizable knowledge from the data they have collected (e.g., how much better one teaching strategy is compared to another) for the sake of scientific discovery and informed decision making.

In this chapter we will focus on methods to construct confidence intervals for the margin—the difference in expected rewards of two bandit arms—from batched bandit data. Rather than constructing high probability confidence intervals, we are interested in constructing confidence intervals by using the asymptotic distribution of estimators to approximate their finite sample distribu-

tion. Asymptotic approximation methods for statistical inference has a long history of being successful in science and leads to much narrower confidence intervals than those constructed using high probability bounds. Most statistical inference methods based on asymptotic approximation assume that treatments are assigned independently[48]. **However, bandit data violates this independence assumption because it is collected *adaptively*, meaning previous actions and rewards inform future action selections.** The non-independence makes statistical inference more challenging, e.g., estimators like the sample mean are often biased on bandit data[76,89].

Throughout, we focus on the batched bandit setting, in which arms of the bandit are pulled in batches. For our asymptotic analysis we fix the total number of batches, T, and allow the arm pulls in each batch, n, to go to infinity. *Note that we do not need or expect n to go to infinity for real-world experiments; we use the asymptotic distribution of estimators to approximate their finite-sample distribution when constructing confidence intervals.* We focus on the batched setting because it closely reflects many of the problem settings where bandit algorithms are applied. For example, in many mobile health[105,61,67] and online education problems[59,82] multiple users use apps / take courses simultaneously, so a batch corresponds to the number of unique users the bandit algorithm acts on at once. The batched setting is even common in online recommendations and advertising because it is impractical to update the bandit after every action if many users visit the site simultaneously[93,87,41,65]. In many such experimental settings the length of the study, T, cannot be arbitrarily adjusted, e.g., in online education, courses generally cannot be made arbitrarily long, and clinical trials often run for a standard amount of time that depends on the domain science (e.g. the length of mobile health studies is a function of the scientific community's belief in how long it should take for users to form a habit). On the other hand, the number of users, n, can in principle grow as large as funding allows.

Additionally, in our batched setting, we assume that the means of the arms can change over time, i.e., from batch to batch, which reflects the temporal non-stationarity that is prevalent in many real

world bandit application problems. For example, in online recommendation systems, the click through rate of a given recommendation typically varies over time, e.g., breaking news articles become less popular over time[93,65]. Online education and mobile health are also highly non-stationary problems because users tend to disengage over time, so the same notification may be much less effective if sent near the end of an experiment than sent near the beginning[33,58,27]. Our statistical inference method does not need to assume that the number of stationary time periods in the experiment is large and is robust to temporal non-stationarity from batch to batch.

The first contribution of this work is proving that on bandit data, rather surprisingly, whether standard estimators are asymptotically normal can depend on whether the margin is zero. We prove that for common bandit algorithms, the arm selection probabilities only concentrate if there is a unique optimal arm. Thus, for two-arm bandits, the arm selection probabilities do not concentrate when the margin—the difference in the expected rewards between the arms—is zero. We show that this leads the ordinary least squares (OLS) estimator to be asymptotically normal when the margin is non-zero, and asymptotically *not* normal when the margin is zero. *Since the OLS estimator does not converge uniformly (over values of the margin), standard inference methods (normal approximations, bootstrap*) can lead to inflated Type-1 error and unreliable confidence intervals on bandit data.*

The second contribution of this work is introducing the Batched OLS (BOLS) estimator, which can be used for reliable inference—even in non-stationary settings—on data collected with batched bandits. We prove that, regardless of whether the margin is zero or not, the BOLS estimator for the margin for both multi-arm and contextual bandits is asymptotically normal and thus can be used for both hypothesis testing and obtaining confidence intervals. Moreover, BOLS is also automatically robust to non-stationarity in the rewards and can be used for constructing valid confidence intervals even if there is non-stationarity in the baseline reward, i.e., if the rewards of

*Note that the validity of bootstrap methods rely on uniform convergence[86].

the arms change from batch to batch, but the margin remains constant. If the margin itself is also non-stationary, BOLS can also be used for constructing simultaneous confidence intervals for the margins for each batch.

3.2 Related Work

Batched Bandits Much work on batched bandits focuses on minimizing regret[78,37] or identifying the best arm with high probability[3,52]. The best arm identification literature utilizes high probability confidence bounds to construct confidence intervals for bandit parameters; we will discuss this method in the next section. Note that in contrast to other batched bandit literature that allow batch sizes to be adjusted adaptively[78], here we do not have adaptive control over the batch sizes.

Batched bandits are closely related to multistage adaptive clinical trials, in which between each batch (or stage of the trial) the data collection procedure can be adjusted depending on the outcome of the previous batches. Our Batched OLS estimator is most closely related to "stage-wise" p-values for group sequential trials that are computed on each stage separately[103]. p-value combination tests are commonly used to combine stage-wise p-values, when the sequence of p-values are shown to be independent or *p-clud*, meaning that under the null each p-value has a Uniform(0,1) distribution conditional on past p-values[103]. Liu et al.[69] formally establish the independence of stage-wise p-values for two-stage trials in which there are a countable number of adaptive rules; note that this rules out bandit algorithms with real-valued arm selection probabilities, like Thompson Sampling.[14] establishes the p-clud property for two-stage adaptive clinical trials under the assumption that the distribution of the second stage data is known conditioned on the decision rule and first stage data under the null hypothesis. Neither of these methods are sufficient for obtaining independent p-values for adaptive trials (1) with an arbitrary number of stages, (2) where exact distribution of rewards is unknown, and (3) where the action selection probabilities can be real numbers,

like for Thompson Sampling.

High Probability Confidence Intervals High probability confidence intervals provide stronger guarantees than those constructed using asymptotic approximations. In particular, these bounds are guaranteed to hold for finite number of observations and often even hold uniformly over all n and T. These types of bounds are used throughout the bandit and reinforcement learning literature to construct confidence intervals for bandit parameters[43,57], prove regret bounds[2,63], and provide guarantees regarding best arm identification[49,50]. The primary drawback of high probability confidence intervals is that they are much more conservative than those constructed using asymptotic approximations. This means that many more observations will be needed to get a confidence interval of the same width or for a statistical test to have the same power when using high probability confidence intervals compared to those constructed using asymptotic approximation. Since the cost of increasing the the number of users in a study can be large, being able to construct narrow—yet reliable—confidence intervals is crucial to many applications.

In our simulations we compare our method to high probability confidence bounds constructed using the self-normalized martingale bound of[2]. This bound is guaranteed to hold on adaptively collected data and is commonly used in the proof of regret bounds for bandit algorithms. We find that all the approaches based on asymptotic approximations (which we discuss next), significantly outperform the statistical test constructed using a self-normalized martingale bound in terms of power. Moreover, despite the weaker guarantees of statistical inference based on asymptotic approximations, they are generally able to provide reliable coverage of confidence intervals and type-1 error control.

Adaptive Inference based on Asymptotic Approximations A common approach in the literature for performing inference on bandit data is to use adaptive weights, which are weights that are a function of the history. An early example of using adaptive weights is that of Luedtke & Van Der Laan[72] and Luedtke & Laan[71], who use adaptive weights in estimating the expected reward

under the optimal policy when one has access to i.i.d. observational data. They use an Augmented-Inverse-Probability-Weighted estimator with adaptive weights that are a function of the estimated standard deviation of the reward. Luedtke & Laan [71] conjecture that their approach can be adapted to the adaptive sampling case. Subsequently Hadad et al. [39] developed the adaptively weighted method for inference on bandit data to produce the Adaptively-Weighted Augmented-Inverse-Probability-Weighted Estimator (AW-AIPW) for data collected via multi-arm bandits. They prove a central limit theorem (CLT) for AW-AIPW when the adaptive weights satisfy certain conditions. Note, however, the AW-AIPW estimator does not have guarantees in non-stationary settings.

Adaptive weights are also used by Deshpande et al. [26] to form the W-decorrelated estimator, a debiased version of OLS, that is asymptotically normal. In the multi-arm bandit setting, the adaptive weights are a function of the number of times an arm was chosen previously. We found that in the two-arm setting, the W-decorrelated estimator down-weights rewards from later in the study (Appendix A.6). Deshpande et al. [25] introduce the Online Debiased Estimator that also has bias guarantees on adaptive data, but in the more challenging high-dimensional linear regression setting. They prove the asymptotic normality of their estimator in the Gaussian autoregressive time series and the two-batch settings. Note that none of these estimation methods have guarantees in non-stationary bandit settings.

Lai & Wei [62] provide conditions under which the OLS estimator is asymptotically normal on adaptively collected data. However, as noted in Villar et al. [101], Deshpande et al. [26], Hadad et al. [39], classical inference techniques developed for i.i.d. data often empirically have inflated Type-1 error on bandit data. In Section 3.4.1, we discuss the restrictive nature of the CLT conditions of Lai & Wei [62].

3.3 Problem Formulation

Setup and Notation Though our results generalize to K-arm, contextual bandits (see Section 3.5.2), we first focus on the two-arm bandit for expositional simplicity. Suppose there are T timesteps or batches in a study. In each batch $t \in [1\colon T]$, we select n binary actions $\{A_{t,i}\}_{i=1}^n \in \{0,1\}^n$. We then observe independent rewards $\{R_{t,i}\}_{i=1}^n$, one for each action selected. Note that the distribution of these random variables changes with the batch size, n. For example, the distribution of the actions one chooses for the 2nd batch, $\{A_{2,i}\}_{i=1}^n$, may change if one has observed $n = 10$ vs. $n = 100$ samples $\{A_{1,i}, R_{1,i}\}_{i=1}^n$ in the first batch. For readability, we omit indexing random variables by n, except for the variables $H_{t-1}^{(n)}$ and $\pi_t^{(n)}$, and filtrations like $\mathcal{G}_{t-1}^{(n)}$ to be introduced next.

For each $t \in [1\colon T]$, the bandit selects actions $\{A_{t,i}\}_{i=1}^n \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi_t^{(n)})$ conditional on $H_{t-1}^{(n)} := \{A_{t',i}, R_{t',i}\}_{i=1,t'=1}^{i=n,t'=t-1}$, the history prior to batch t. Note, the *action selection probability* $\pi_t^{(n)} := \mathbb{P}(A_{t,i} = 1 | H_{t-1}^{(n)})$ depends on the history $H_{t-1}^{(n)}$. We assume the following conditional mean for rewards:

$$\mathbb{E}\left[R_{t,i} \middle| H_{t-1}^{(n)}, A_{t,i}\right] = (1 - A_{t,i})\beta_{t,0} + A_{t,i}\beta_{t,1}. \tag{3.3.1}$$

Note in equation (3.3.1) we condition on $H_{t-1}^{(n)}$ because the conditional mean of the reward does not depend on prior rewards or actions. Let $\mathbf{X}_{t,i} := [1 - A_{t,i}, A_{t,i}]^\top \in \mathbb{R}^2$; note $\mathbf{X}_{t,i}$ is higher dimensional when we add more arms and/or context variables. We define the errors as $\varepsilon_{t,i} := R_{t,i} - (\mathbf{X}_{t,i})^\top \beta_t$. Equation (3.3.1) implies that $\{\varepsilon_{t,i} : i \in [1\colon n], t \in [1\colon T]\}$ are a martingale difference array with respect to the filtration $\{\mathcal{G}_t^{(n)}\}_{t=1}^T$, where $\mathcal{G}_t^{(n)} := \sigma\big(H_{t-1}^{(n)} \cup \{A_{t,i}\}_{i=1}^n\big)$; thus, $\mathbb{E}[\varepsilon_{t,i} | \mathcal{G}_{t-1}^{(n)}] = 0, \forall t, i, n$. The parameters $\beta_t = (\beta_{t,0}, \beta_{t,1})$ can change across batches $t \in [1\colon T]$, which allows for non-stationarity between batches. Assuming that $\beta_t = \beta_{t'}$ for all $t, t' \in [1\colon T]$ simplifies to the stationary mean case.

Action Selection Probability Constraint (Clipping) In order to perform inference on bandit data it is necessary to guarantee that the bandit algorithm explores sufficiently. For example, the CLTs for both the W-decorrelated [26] and the AW-AIPW [39] estimators have conditions that implicitly require that the bandit algorithms cannot sample any given action with probability that goes to zero or one arbitrarily fast. Greater exploration also increases the power of statistical tests regarding the margin [104]. Moreover, if there is non-stationarity in the margin between batches, it is desirable for the bandit algorithm to continue exploring. We explicitly guarantee exploration by constraining the probability that any given action can be sampled (see Definition 3.3.1). We allow the action selection probabilities $\pi_t^{(n)}$ to converge to 0 and/or 1 at some rate.

Definition 3.3.1. *A clipping constraint with rate $f(n)$ means that $\pi_t^{(n)}$ satisfies the following:*

$$\lim_{n\to\infty} \mathbb{P}\big(\pi_t^{(n)} \in [f(n), 1-f(n)]\big) = 1 \tag{3.3.2}$$

3.4 Asymptotic Distribution of the Ordinary Least Squares Estimator

Suppose we are in the stationary case, and we would like to estimate β. Consider the OLS estimator: $\hat{\beta}^{\text{OLS}} = (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{R}$, where $\underline{\mathbf{X}} := [\mathbf{X}_{1,1}, .., \mathbf{X}_{1,n}, .., \mathbf{X}_{T,1}, .., \mathbf{X}_{T,n}]^\top \in \mathbb{R}^{nT\times 2}$ and $\mathbf{R} := [R_{1,1}, .., R_{1,n}, .., R_{T,1}, .., R_{T,n}]^\top \in \mathbb{R}^{nT}$. Note that $\underline{\mathbf{X}}^\top \underline{\mathbf{X}} = \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{t,i} \mathbf{X}_{t,i}^\top$.

3.4.1 Conditions for Asymptotically Normality of the OLS estimator

If $(\mathbf{X}_{t,i}, \varepsilon_{t,i})$ are i.i.d., $\mathbb{E}[\varepsilon_{t,i}] = 0$, $\mathbb{E}[\varepsilon_{t,i}^2] = \sigma^2$, and the first two moments of $\mathbf{X}_{t,i}$ exist, a classical result from statistics [6] is that the OLS estimator is asymptotically normal, i.e., as $n \to \infty$,

$$(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{1/2} (\hat{\beta}^{\text{OLS}} - \beta) \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{I}_p).$$

[62] generalize this result by proving that the OLS estimator is still asymptotically normal in the adaptive sampling case when $\underline{\mathbf{X}}^\top \underline{\mathbf{X}}$ satisfies a certain stability condition. To show that a similar result holds for the batched setting, we generalize the CLT of Lai & Wei [62] to triangular arrays (required since the distribution of our random variables vary as the batch size, n, changes), as stated in Theorem 3.4.1.

Condition 3.4.1 (Moments). *For all* t, n, i, $\mathbb{E}[\varepsilon_{t,i}^2 \big| \mathcal{G}_{t-1}^{(n)}] = \sigma^2$ *and* $\mathbb{E}\left[\varepsilon_{t,i}^4 \big| \mathcal{G}_{t-1}^{(n)}\right] < M < \infty$.

Condition 3.4.2 (Stability). *For some non-random sequence of scalars* $\{a_i\}_{i=1}^{\infty}$, *as* $n \to \infty$,

$$a_n \cdot \tfrac{1}{nT} \textstyle\sum_{t=1}^{T} \sum_{i=1}^{n} A_{t,i} \xrightarrow{P} 1$$

Theorem 3.4.1 (Triangular array version of Lai & Wei [62], Theorem 3). *Assuming Conditions 3.4.1 and 3.4.2, as* $n \to \infty$,

$$\left(\underline{\boldsymbol{X}}^\top \underline{\boldsymbol{X}}\right)^{1/2} (\hat{\beta}^{\mathrm{OLS}} - \beta) \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_p).$$

Note that in the bandit setting, Condition 3.4.2 means that prior to running the experiment, the asymptotic rate at which arms will be selected is predictable. We will show that Condition 3.4.2 is in a sense necessary for the asymptotic normality of OLS. In Corollary 3.4.1 below we state that Conditions 3.4.1 and 3.4.3, and a non-zero margin are sufficient for stability Condition 3.4.2. Later, we will show that when the margin is zero, Condition 3.4.2 does not hold for many common bandit algorithms and prove that this leads the OLS estimator to be asymptotically non-normal.

Condition 3.4.3 (Conditionally i.i.d. actions). *For each* $t \in [1\colon T]$, $\{A_{t,i}\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathrm{Bernoulli}(\pi_t^{(n)})$ *i.i.d. over* $i \in [1\colon n]$ *conditional on* $H_{t-1}^{(n)}$.

Corollary 3.4.1 (Sufficient conditions for Theorem 3.4.1). *If Conditions 3.4.1 and 3.4.3 hold, and* ***the margin is non-zero****, data collected in batches using* ε*-greedy, Thompson Sampling, or UCB with clipping constraint with* $f(n) = c$ *for some* $0 < c \leq \frac{1}{2}$ *(see Definition 3.3.1) satisfy Theorem 3.4.1 conditions.*

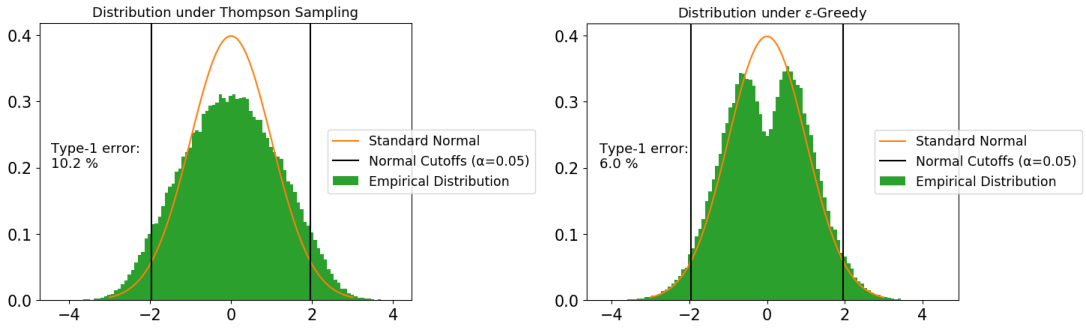

Figure 3.1: Empirical distribution of the Z-statistic (σ^2 is known) of the OLS estimator for the margin. All simulations are with no margin ($\beta_1 = \beta_0 = 0$); $\mathcal{N}(0,1)$ rewards; $T = 25$; and $n = 100$. For ε-greedy, $\varepsilon = 0.1$.

3.4.2 Asymptotic Non-Normality under No Margin

We prove the conjecture of[26] that when the margin is zero, the OLS estimator is asymptotically non-normal under common bandit algorithms, including Thompson Sampling, ε-greedy, and UCB. Thus as seen in Figure 3.1, assuming the OLS estimator is approximately Normal on bandit data can lead to inflated Type-1 error, even asymptotically. The asymptotic non-normality of OLS occurs when the margin is zero because when there is no unique optimal arm, $\pi_t^{(n)}$ does not concentrate as $n \to \infty$ (Appendix A.3).

We state the asymptotic non-normality result for Thompson Sampling in Theorem 3.4.2; see Appendix A.3 for the proof and similar results for ε-greedy and UCB. It is sufficient to prove asymptotic non-normality for $T = 2$. Note, $\hat{\Delta}^{\text{OLS}}$ is the difference in the sample means for each arm, so $\hat{\Delta}^{\text{OLS}} = \hat{\beta}_1^{\text{OLS}} - \hat{\beta}_0^{\text{OLS}}$. The Z-statistic of $\hat{\Delta}^{\text{OLS}}$, which is asymptotically normal under i.i.d. sampling, is as follows:

$$\sqrt{\frac{(\sum_{t=1}^{2}\sum_{i=1}^{n} A_{t,i})(\sum_{t=1}^{2}\sum_{i=1}^{n} 1 - A_{t,i})}{2\sigma^2 n}}\left(\hat{\Delta}^{\text{OLS}} - \Delta\right). \tag{3.4.1}$$

Theorem 3.4.2 (Asymptotic non-normality of OLS estimator under zero margin for Thompson Sampling)**.** *Let $T = 2$ and $\pi_1^{(n)} = \frac{1}{2}$. If $\varepsilon_{t,i} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, we have independent normal priors on*

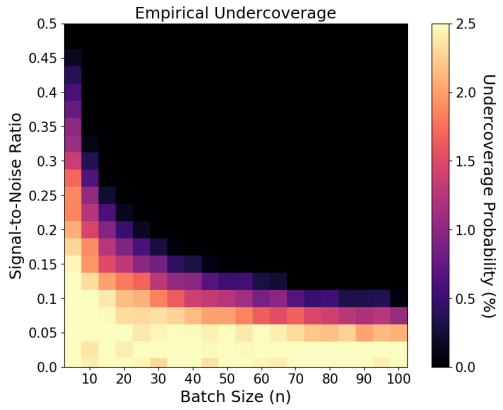

Figure 3.2: Empirical undercoverage probabilities (coverage probability below 95%) of confidence intervals using on a normal approximation for the OLS estimator. We use Thompson Sampling with $\mathcal{N}(0,1)$ priors, a clipping constraint of $0.05 \leq \pi_t^{(n)} \leq 0.95$, $\mathcal{N}(0,1)$ rewards, $T = 25$, and known σ^2. Standard errors are < 0.001.

arm means $\tilde{\beta}_0, \tilde{\beta}_1 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, *and* $\pi_2^{(n)} = \pi_{\min} \vee [(1-\pi_{\max}) \wedge \mathbb{P}(\tilde{\beta}_1 > \tilde{\beta}_0 | H_1^{(n)})]$ *for constants* $\pi_{\min}, \pi_{\max}$ *with* $0 < \pi_{\min} \leq \pi_{\max} < 1$, *then* (3.4.1) *is asymptotically* ***not*** *normal when the margin is zero.*

Since the OLS estimator is asymptotically normal when $\Delta \neq 0$ (Corollary 3.4.1) and asymptotically *not* Normal when $\Delta = 0$, the OLS estimator does not converge *uniformly* on data collected under standard bandit algorithms. The non-uniform convergence of the OLS estimator precludes us from using a normal approximation to perform hypothesis testing and construct confidence intervals (see [55]). In real-world applications, there is rarely exactly zero margin. However, the non-uniform convergence of the OLS estimator at zero margin is still practically important because the asymptotic distribution of the OLS estimator when the margin is zero is indicative of the finite-sample distribution when the margin is statistically difficult to differentiate from zero, i.e., when the signal-to-noise ratio, $\frac{|\Delta|}{\sigma}$, is low. Figure 3.2 shows that *even when the margin is non-zero, when the signal-to-noise ratio is low, confidence intervals constructed using a normal approximation have coverage probabilities below the nominal level. Moreover, for any batch size n and noise variance* σ^2*, there exists a non-zero margin size with a finite-sample distribution that is poorly approximated by a*

normal distribution.

3.5 Batched OLS Estimator

3.5.1 Batched OLS Estimator for Multi-Arm Bandits

We now introduce the Batched OLS (BOLS) estimator that is asymptotically normal under a large class of bandit algorithms, even when the margin is zero. Instead of computing the OLS estimator on all the data, we compute the OLS estimator for each batch and normalize it by the variance estimated from that batch. For each $t \in [1\colon T]$, the BOLS estimator of the margin $\Delta_t := \beta_{t,1} - \beta_{t,0}$ is:

$$\hat{\Delta}_t^{\text{BOLS}} = \frac{\sum_{i=1}^n (1 - A_{t,i}) R_{t,i}}{\sum_{i=1}^n 1 - A_{t,i}} - \frac{\sum_{i=1}^n A_{t,i} R_{t,i}}{\sum_{i=1}^n A_{t,i}}.$$

Theorem 3.5.1 (Asymptotic Normality of Batched OLS estimator for multi-arm bandits)**.** *Assuming Conditions 3.4.1 (moments) and 3.4.3 (conditionally i.i.d. actions), and a clipping rate $f(n) = \frac{1}{n^\alpha}$ for some $0 \leq \alpha < 1$ (see Definition 3.3.1),*

$$\begin{bmatrix} \sqrt{\frac{(\sum_{i=1}^n 1 - A_{1,i})(\sum_{i=1}^n A_{1,i})}{n}} (\hat{\Delta}_1^{\text{BOLS}} - \Delta_1) \\ \sqrt{\frac{(\sum_{i=1}^n 1 - A_{2,i})(\sum_{i=1}^n A_{2,i})}{n}} (\hat{\Delta}_2^{\text{BOLS}} - \Delta_2) \\ \vdots \\ \sqrt{\frac{(\sum_{i=1}^n 1 - A_{T,i})(\sum_{i=1}^n A_{T,i})}{n}} (\hat{\Delta}_T^{\text{BOLS}} - \Delta_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_T)$$

It is straightforward to generalize Theorem 3.5.1 to the case that batches are different sizes but the size of the smallest batch goes to infinity and the batch size is independent of the history.

By Theorem 3.5.1, for the stationary margin case, we can test $H_0 : \Delta = c$ vs. $H_1 : \Delta \neq c$ with the

following statistic, which is asymptotically normal under the null:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sqrt{\frac{(\sum_{i=1}^{n}1-A_{t,i})(\sum_{i=1}^{n}A_{t,i})}{n\sigma^2}}(\hat{\Delta}_t^{\text{BOLS}}-c). \tag{3.5.1}$$

This type of test statistic—a weighted combination of asymptotically independent normals—a special case of the inverse normal p-value combination test, has been used in simple settings in which the studies (e.g., batches) are independent (e.g., when conducting meta-analyses across multiple studies) [26]. Here the ability to use this type of test statistic is novel since, due to the bandit algorithm, the batches are *not* independent. The work here demonstrates asymptotic independence and thus for large n the Z-statistics from each batch should be approximately independently distributed.

The key to proving asymptotic normality for BOLS is that the following ratio converges in probability to one: $\frac{(\sum_{i=1}^{n}1-A_{t,i})(\sum_{i=1}^{n}A_{t,i})}{n}\frac{1}{n\pi_t^{(n)}(1-\pi_t^{(n)})}\xrightarrow{P}1$. Since $\pi_t^{(n)}\in\mathcal{G}_{t-1}^{(n)}$, $\frac{1}{n\pi_t^{(n)}(1-\pi_t^{(n)})}$ is a constant given $\mathcal{G}_{t-1}^{(n)}$. Thus, even if $\pi_t^{(n)}$ does not concentrate, we are still able to apply the martingale CLT[29] to prove asymptotic normality. See Appendix A.2 for more details.

3.5.2 Batched OLS Estimator for Contextual Bandits

For contextual K-arm bandits, for any two arms $x,y\in[0\colon K-1]$, we can estimate the margin between them $\Delta_{t,x-y}:=\beta_{t,x}-\beta_{t,y}\in\mathbb{R}^d$. In each batch, we observe context vectors $\{\mathbf{C}_{t,i}\}_{i=1}^{n}$ for $\mathbf{C}_{t,i}\in\mathbb{R}^d$. We redefine the history $H_{t-1}^{(n)}:=\{\mathbf{C}_{t',i},A_{t',i},R_{t',i}\}_{i=1,t'=1}^{i=1,t'=t-1}$ and define the filtration $\mathcal{F}_t^{(n)}:=\sigma\big(H_{t-1}^{(n)}\cup\{A_{t,i},\mathbf{C}_{t,i}\}_{i=1}^{n}\big)$. The action selection probabilities $\pi_t^{(n)}$ are now functions of the context, so $\pi_t^{(n)}(\mathbf{C}_{t,i})\in[0,1]^K$ is a vector where the k^{th} dimension equals $\mathbb{P}(A_{t,i}=k|\mathcal{H}_{t-1}^{(n)},\mathbf{C}_{t,i})$. We assume the following conditional mean model of the reward: $\mathbb{E}\big[R_{t,i}\big|\mathcal{F}_{t-1}^{(n)}\big]=\sum_{k=0}^{K-1}\mathbb{I}_{(A_{t,i}=k)}\mathbf{C}_{t,i}^{\top}\beta_{t,k}$ and let $\varepsilon_{t,i}:=R_{t,i}-\sum_{k=0}^{K-1}\mathbb{I}_{(A_{t,i}=k)}\mathbf{C}_{t,i}^{\top}\beta_{t,k}$.

Condition 3.5.1 (Conditionally i.i.d. contexts)**.** *For each t, $\boldsymbol{C}_{t,1},\boldsymbol{C}_{t,2},\ldots,\boldsymbol{C}_{t,n}$ are i.i.d. and its first two moments, μ_t,Σ_t, are non-random given $H_{t-1}^{(n)}$, i.e., $\mu_t,\Sigma_t\in\sigma(H_{t-1}^{(n)})$.*

Condition 3.5.2 (Bounded context). $||\boldsymbol{C}_{t,i}||_{\max} \leq u$ *for all* i, t, n *for some constant* u. *Also, the minimum eigenvalue of* $\underline{\Sigma}_t$ *is lower bounded, i.e.,* $\lambda_{\min}(\underline{\Sigma}_t) > l > 0$.

Definition 3.5.1. *A conditional clipping constraint with rate* $f(n)$ *means that the action selection probabilities* $\pi_t^{(n)} : \mathbb{R}^d \to [0,1]^K$ *satisfy the following:*

$$\mathbb{P}\big(\forall\, \boldsymbol{c} \in \mathbb{R}^d, \pi_t^{(n)}(\boldsymbol{c}) \in \big[f(n), 1 - f(n)\big]^K\big) \to 1$$

For each $t \in [1\colon T]$, we have the OLS estimator for $\Delta_{t,x-y}$: $\hat{\Delta}_t^{\text{OLS}} := \big[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\big]^{-1}\big(\hat{\beta}_{t,x}^{\text{OLS}} - \hat{\beta}_{t,y}^{\text{OLS}}\big)$, where $\underline{\mathbf{C}}_{t,k} := \sum_{i=1}^n \mathbb{I}_{A_{t,i}^{(n)}=k} \mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \in \mathbb{R}^{d\times d}, \hat{\beta}_{t,k}^{\text{OLS}} = \underline{\mathbf{C}}_{t,k}^{-1} \sum_{i=1}^n \mathbb{I}_{A_{t,i}^{(n)}=k} \mathbf{C}_{t,i} R_{t,i}$.

Theorem 3.5.2 (Asymptotic Normality of Batched OLS estimator for contextual bandits)**.** *Assuming Conditions 3.4.1 (moments)*[†]*, 3.4.3 (conditionally i.i.d. actions), 3.5.1, and 3.5.2, and a conditional clipping rate* $f(n) = c$ *for some* $0 \leq c < \frac{1}{2}$ *(see Definition 3.5.1),*

$$\begin{bmatrix} \big[\underline{\boldsymbol{C}}_{1,x}^{-1} + \underline{\boldsymbol{C}}_{1,y}^{-1}\big]^{1/2}(\hat{\Delta}_1^{\text{OLS}} - \Delta_{1,x-y}) \\ \big[\underline{\boldsymbol{C}}_{2,x}^{-1} + \underline{\boldsymbol{C}}_{2,y}^{-1}\big]^{1/2}(\hat{\Delta}_2^{\text{OLS}} - \Delta_{2,x-y}) \\ \vdots \\ \big[\underline{\boldsymbol{C}}_{T,x}^{-1} + \underline{\boldsymbol{C}}_{T,y}^{-1}\big]^{1/2}(\hat{\Delta}_T^{\text{OLS}} - \Delta_{T,x-y}) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_{Td}).$$

3.5.3 Batched OLS Statistic for Non-Stationary Bandits

Many real-world problems we would like to use bandit algorithms for have non-stationary over time. For example, in online advertising, the effectiveness of an ad may change over time due to exposure to competing ads and general societal changes that could affect perceptions of an ad. We may believe that the expected reward for a given action may vary over time, but that the margin is constant from batch to batch. In the online advertising setting, this would mean whether one ad is always better

[†]Assume an analogous moment condition for the contextual bandit case, where $\mathcal{G}_t^{(n)}$ is replaced by $\mathcal{F}_t^{(n)}$.

than another is stable, but the overall effectiveness of both ads may change over time. In this case, we can simply use the BOLS test statistic described earlier in equation (3.5.1) to test $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$. Note that the BOLS test statistic for the margin is robust to non-stationarity in the baseline reward without any adjustment. Moreover, in our simulation settings we estimate the variance σ^2 separately for each batch, which allows for non-stationarity in the variance between batches as well; see Appendix A.1 for variance estimation details and see Section 3.6 for simulation results. Additionally, in the case that we believe that the margin itself may vary from batch to batch, the BOLS test statistic can also be used to construct confidence regions that contain the true margin Δ_t for each batch simultaneously; see Appendix A.1.5 for details.

3.6 Simulation Experiments

Procedure We focus on the two-arm bandit setting and test whether the margin is zero, specifically $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$. We perform experiments for when the noise variance σ^2 is estimated. We assume homoscedastic errors throughout. See Appendix A.1.4 for more details about how we estimate the noise variance and more details regarding our experimental setup. In Figures A.1 and A.2, we display results for stationary bandits and in Figure 3.5 we show results for bandits with non-stationary baseline rewards. See Appendix A.1.5 for results for bandits with non-stationary margins.

In our simulations, we found that OLS and AW-AIPW have inflated Type-1 error. Since Type-1 control is a hard constraint, solutions with inflated Type-1 error are *infeasible* solutions. In the power plots, we adjust the cutoffs of the estimators to ensure proper Type-1 error control; if an estimator has inflated Type-1 error under the null, in the power simulations we use a critical value estimated using the simulations under the null. Note that it is unfeasible to make these cutoff adjustment for real experiments (unless one found the worst case setting), as there are many nuisance

parameters—like the expected rewards for each arm and the noise variance—which can affect cutoff values.

RESULTS Figure 3.3 shows that for small sample sizes ($nT \lesssim 300$), BOLS has more reliable Type-1 error control than AW-AIPW with variance stabilizing weights. After $nT \geq 500$ samples, AW-AIPW has proper Type-1 error, and by Figure 3.4 it always has slightly greater power than BOLS in the stationary setting. The W-decorrelated estimator has reliable Type-1 error control, but very low power compared to AW-AIPW and BOLS. Finally the high probability, self-normalized martingale bound of[2], which we use for hypothesis testing, has very low power compared to the asymptotic approximation statistical inference methods.

In Figure 3.5, we display simulation results for the non-stationary baseline reward setting. *While other estimators have no Type-1 error guarantees, BOLS still has proper Type-1 error control in the non-stationary baseline reward setting. Moreover, BOLS can have much greater power than other estimators when there is non-stationarity in the baseline reward.* Overall, BOLS is favorable over other estimators in small-sample settings or when one wants to be robust to non-stationarity in the baseline reward—at the cost of losing a little power if the environment is stationary.

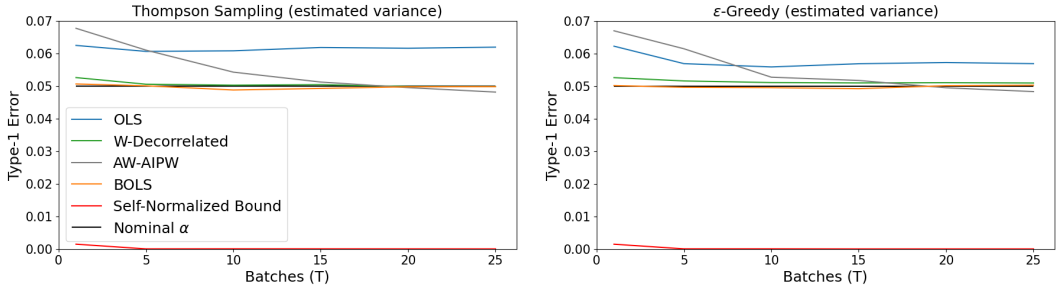

Figure 3.3: Stationary Setting: Type-1 error for a two-sided test of $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ ($\alpha = 0.05$). We set $\beta_1 = \beta_0 = 0, n = 25$, and a clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. We use 100k Monte Carlo simulations and standard errors are < 0.001.

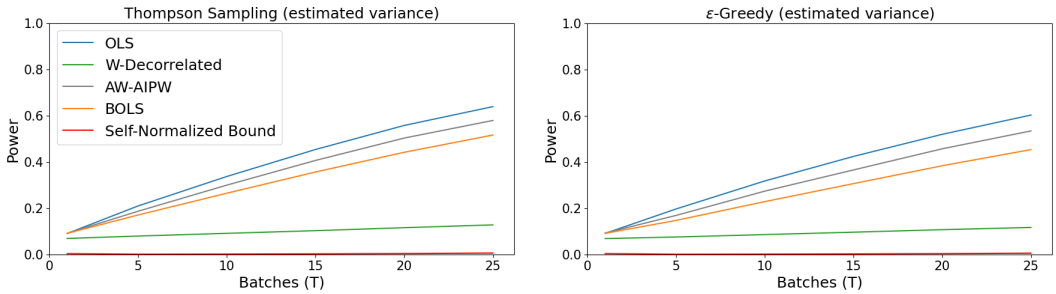

Figure 3.4: Stationary Setting: Power for a two-sided test of $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ ($\alpha = 0.05$). We set $\beta_1 = 0$, $\beta_0 = 0.25, n = 25$, and a clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. We use 100k Monte Carlo simulations and standard errors are < 0.002. We account for Type-1 error inflation as described in Section 3.6.

3.7 Discussion

We found that the OLS estimator is asymptotically non-normal when the margin is zero due to the non-concentration of the action selection probabilities. Since the OLS estimator is a canonical example of a method-of-moments estimator[42], our results suggest that the inferential guarantees of standard method-of-moments estimators may fail to hold on adaptively collected data when there is no unique optimal, regret-minimizing policy. We develop the Batched OLS estimator, which is asymptotically normal even when the action selection probabilities do not concentrate. An open question is whether batched versions of general method-of-moments estimators could similarly be used for adaptive inference.

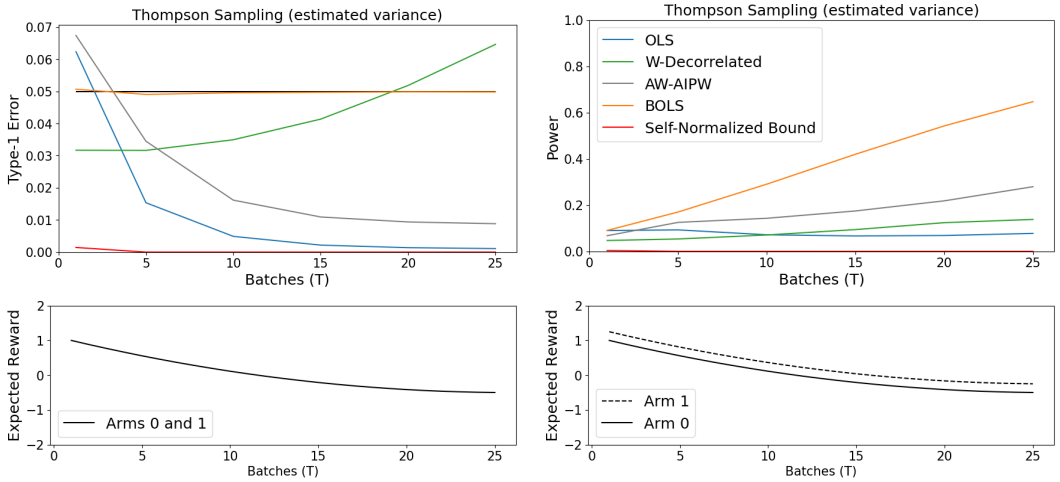

Figure 3.5: Non-stationary baseline reward setting: Type-1 error (upper left) and power (upper right) for a two-sided test of $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ ($\alpha = 0.05$). In the lower two plots we plot the expected rewards for each arm; note the margin is constant across batches. We use $n = 25$ and a clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. We use 100k Monte Carlo simulations and standard errors are < 0.002.

4

Adaptively Weighted M-Estimators

Online reinforcement learning and other adaptive sampling algorithms are increasingly used in digital intervention experiments to optimize treatment delivery for users over time. In this work, we focus on longitudinal user data collected by a large class of adaptive sampling algorithms that are designed to optimize treatment decisions online using accruing data from multiple users. Combining or "pooling" data across users allows adaptive sampling algorithms to

potentially learn faster. However, by pooling, these algorithms induce dependence between the sampled user data trajectories; we show that this can cause standard variance estimators for i.i.d. data to underestimate the true variance of common estimators on this data type. We develop novel methods to perform a variety of statistical analyses on such adaptively sampled data via Z-estimation. Specifically, we introduce the *adaptive* sandwich variance estimator, a corrected sandwich estimator that leads to consistent variance estimates under adaptive sampling. Additionally, to prove our results we develop novel theoretical tools for empirical processes on non-i.i.d., adaptively sampled longitudinal data which may be of independent interest. This work is motivated by our efforts in designing experiments in which online reinforcement learning algorithms optimize treatment decisions, yet statistical inference is essential for conducting analyses after experiments conclude.

4.1 Introduction

Due to the need for interventions that are personalized to users, (contextual) bandit algorithms are increasingly used to address sequential decision making problems in health-care [105,67], online education [70,88], and public policy [56,19]. Contextual bandits personalize, that is, minimize regret, by learning to choose the best intervention in each context, i.e., the action that leads to the greatest expected reward. Besides the goal of regret minimization, another critical goal in these real-world problems is to be able to use the resulting data collected by bandit algorithms to advance scientific knowledge [70,32]. By scientific knowledge, we mean information gained by using the data to conduct a variety of statistical analyses, including confidence interval construction and hypothesis testing. **While regret minimization is a *within*-experiment learning objective, gaining scientific knowledge from the resulting adaptively collected data is a *between*-experiment learning objective**, which ultimately helps with regret minimization between deployments of bandit algorithms. Note that the data collected by bandit algorithms are *adaptively collected* because previously observed contexts,

actions, and rewards are used to inform what actions to select in future timesteps.

There are a variety of between-experiment learning questions encountered in real-life applications of bandit algorithms. For example, in real-life sequential decision-making problems there are often a number of additional scientifically interesting outcomes besides the reward that are collected during the experiment. In the online advertising setting, the reward might be whether an ad is clicked on, but one may be interested in the outcome of amount of money spent or the subsequent time spent on the advertiser's website. If it was found that an ad had high click-through rate, but low amounts of money was spent after clicking on the ad, one may redesign the reward used in the next bandit experiment. One type of statistical analysis would be to construct confidence intervals for the relative effect of the actions on multiple outcomes (in addition to the reward) conditional on the context. Furthermore, due to engineering and practical limitations, some of the variables that might be useful as context are often not accessible to the bandit algorithm online. If after-study analyses find some such contextual variables to have sufficiently strong influence on the relative usefulness of an action, this might lead investigators to ensure these variables are accessible to the bandit algorithm in the next experiment.

As discussed above, we can gain scientific knowledge from data collected with (contextual) bandit algorithms by constructing confidence intervals and performing hypothesis tests for unknown quantities such as the expected outcome for different actions in various contexts. Unfortunately, standard statistical methods developed for i.i.d. data fail to provide valid inference when applied to data collected with common bandit algorithms. For example, assuming the sample mean of rewards for an arm is approximately normal can lead to unreliable confidence intervals and inflated type-1 error; see Section 4.3.1 for an illustration. Recently statistical inference methods have been developed for data collected using bandit algorithms [39,26,109]; however, these methods are limited to inference for parameters of simple models. There is a lack of general statistical inference methods for data collected with (contextual) bandit algorithms in more complex data-analytic settings, includ-

ing parameters in non-linear models for outcomes; for example, there are currently no methods for constructing valid confidence intervals for the parameters of a logistic regression model for binary outcomes or for constructing confidence intervals based on robust estimators like minimizers of the Huber loss function.

In this work we show that a wide variety of estimators which are frequently used both in science and industry on i.i.d. data, namely, M-estimators[99], can be used to conduct valid inference on data collected with (contextual) bandit algorithms when adjusted with particular adaptive weights, i.e., weights that are a function of previously collected data. Different forms of adaptive weights are used by existing methods for simple models[26,39,109]. Our work is a step towards developing a general framework for statistical inference on data collected with adaptive algorithms, including (contextual) bandit algorithms.

4.2 Problem Formulation

We assume that the data we have after running a contextual bandit algorithm is comprised of contexts $\{X_t\}_{t=1}^T$, actions $\{A_t\}_{t=1}^T$, and primary outcomes $\{Y_t\}_{t=1}^T$. T is deterministic and known. We assume that rewards are a deterministic function of the primary outcomes, i.e., $R_t = f(Y_t)$ for some known function f. We are interested in constructing confidence regions for the parameters of the conditional distribution of Y_t given (X_t, A_t). Below we consider $T \to \infty$ in order to derive the asymptotic distributions of estimators and construct asymptotically valid confidence intervals. We allow the action space $\mathcal{A}$ to be finite or infinite. We use potential outcome notation[48] and let $\{Y_t(a) : a \in \mathcal{A}\}$ denote the potential outcomes of the primary outcome and let $Y_t \triangleq Y_t(A_t)$ be the observed outcome. We assume a stochastic contextual bandit environment in which $\{X_t, Y_t(a) : a \in \mathcal{A}\} \overset{i.i.d.}{\sim} \mathcal{P} \in \mathbf{P}$ for $t \in [1 : T]$; the contextual bandit environment distribution $\mathcal{P}$ is in a space of possible environment distributions $\mathbf{P}$. We define the history

$\mathcal{H}_t \triangleq \{X_{t'}, A_{t'}, Y_{t'}\}_{t'=1}^{t}$ for $t \geq 1$ and $\mathcal{H}_0 \triangleq \emptyset$. Actions $A_t \in \mathcal{A}$ are selected according to policies $\pi \triangleq \{\pi_t\}_{t\geq 1}$, which define action selection probabilities $\pi_t(A_t, X_t, \mathcal{H}_{t-1}) \triangleq \mathbb{P}(A_t|\mathcal{H}_{t-1}, X_t)$. Even though the potential outcomes are i.i.d., the *observed* data $\{X_t, A_t, Y_t\}_{t=1}^{T}$ are *not* because the actions are selected using policies π_t which are a function of past data, $\mathcal{H}_{t-1}$. Non-independence of observations is a key property of adaptively collected data.

We are interested in constructing confidence regions for some unknown $\theta^*(\mathcal{P}) \in \Theta \subset \mathbb{R}^d$, which is a parameter of the conditional distribution of Y_t given (X_t, A_t). This work focuses on the setting in which we have a well-specified model for Y_t. Specifically, we assume that $\theta^*(\mathcal{P})$ is a conditionally maximizing value of criterion m_θ, i.e., for all $\mathcal{P} \in \mathbf{P}$,

$$\theta^*(\mathcal{P}) \in \operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}_{\mathcal{P}}\left[m_\theta(Y_t, X_t, A_t)|X_t, A_t\right] \text{ w.p. 1.} \tag{4.2.1}$$

Note that $\theta^*(\mathcal{P})$ does not depend on (X_t, A_t) and it is an implicit modelling assumption that such a $\theta^*(\mathcal{P})$ exists for a given m_θ. Note that this formulation includes semi-parametric models, e.g., the model could constrain the conditional mean of Y_t to be linear in some function of the actions and context, but allow the residuals to follow any mean-zero distribution, including ones that depend on the actions and/or contexts.

To estimate $\theta^*(\mathcal{P})$, we build on M-estimation [46], which classically selects the estimator $\hat{\theta}$ to be the $\theta \in \Theta$ that maximizes the empirical analogue of Equation (4.2.1):

$$\hat{\theta}_T \triangleq \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} m_\theta(Y_t, X_t, A_t). \tag{4.2.2}$$

For example, in a classical linear regression setting with $|\mathcal{A}| < \infty$ actions, a natural choice for m_θ is the negative of the squared loss function, $m_\theta(Y_t, X_t, A_t) = -(Y_t - X_t^\top \theta_{A_t})^2$. When Y_t is binary, a natural choice is instead the negative log-likelihood function for a logistic regression model, i.e.,

$m_\theta(Y_t, X_t, A_t) = -[Y_t X_t^\top \theta_{A_t} - \log(1 + \exp(X_t^\top \theta_{A_t}))]$. More generally, m_θ is commonly chosen to be a log-likelihood function or the negative of a robust loss function such as the Huber loss. If the data, $\{X_t, A_t, Y_t\}_{t=1}^T$, were independent across time, classical approaches could be used to prove the consistency and asymptotic normality of M-estimators[99]. However, on data collected with bandit algorithms, standard M-estimators like the ordinary least-squares estimator fail to provide valid confidence intervals[39,26,109]. In this work, we show that M-estimators can still be used to provide valid statistical inference on adaptively collected data when adjusted with well-chosen adaptive weights.

4.3 Adaptively Weighted M-Estimators

We consider a weighted M-estimating criteria with adaptive weights $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ given by $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$. Here $\{\pi_t^{\text{sta}}\}_{t \geq 1}$ are pre-specified *stabilizing policies* that do not depend on data $\{Y_t, X_t, A_t\}_{t \geq 1}$. A default choice for the stabilizing policy when the action space is of size $|\mathcal{A}| < \infty$ is just $\pi_t^{\text{sta}}(a, x) = 1/|\mathcal{A}|$ for all x, a, and t; we discuss considerations for the choice of $\{\pi_t^{\text{sta}}\}_{t=1}^T$ in Section 4.3.3. We call these weights *square-root importance weights* because they are the square-root of the standard importance weights[40,102]. Our proposed estimator for $\theta^*(\mathcal{P})$, $\hat{\theta}_T$, is the maximizer of a *weighted* version of the M-estimation criterion of Equation (4.2.2):

$$\hat{\theta}_T \triangleq \operatorname*{argmax}_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} W_t m_\theta(Y_t, X_t, A_t) =: \operatorname*{argmax}_{\theta \in \Theta} M_T(\theta).$$

Note that $M_T(\theta)$ defined above depends on both the data $\{X_t, A_t, Y_t\}_{t=1}^T$ and weights $\{W_t\}_{t=1}^T$. We provide asymptotically valid confidence regions for $\theta^*(\mathcal{P})$ by deriving the asymptotic distribution of $\hat{\theta}_T$ as $T \to \infty$ and by proving that the convergence in distribution is *uniform* over $\mathcal{P} \in \mathbf{P}$. Such convergence allows us to construct a uniformly asymptotically valid $1 - \alpha$ level confidence region,

$C_T(\alpha)$, for $\theta^*(\mathcal{P})$, which is a confidence region that satisfies

$$\liminf_{T\to\infty} \inf_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\theta^*(\mathcal{P}) \in C_T(\alpha)\right) \geq 1-\alpha. \tag{4.3.1}$$

If $C_T(\alpha)$ were *not* uniformly valid, then there would exist an $\varepsilon > 0$ such that for *every* sample size T, $C_T(\alpha)$'s coverage would be below $1-\alpha-\varepsilon$ for some worst-case $P_T \in \mathbf{P}$. Confidence regions which are asymptotically valid, but not *uniformly* asymptotically valid, fail to be reliable in practice[64,86]. Note that on i.i.d. data it is generally straightforward to show that estimators that converge in distribution do so uniformly; however, as discussed in Zhang et al. [109] and Appendix B.4, this is not the case on data collected with bandit algorithms.

To construct uniformly valid confidence regions for $\theta^*(\mathcal{P})$ we prove that $\hat{\theta}_T$ is uniformly asymptotically normal in the following sense:

$$\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}\left(0, I_d\right) \text{ uniformly over } \mathcal{P}\in\mathbf{P}, \tag{4.3.2}$$

where $\ddot{M}_T(\theta) \triangleq \frac{\partial^2}{\partial^2\theta} M_T(\theta)$ and $\Sigma_T(\mathcal{P}) \triangleq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)^{\otimes 2}\right]$. We define $\dot{m}_\theta \triangleq \frac{\partial}{\partial\theta} m_\theta$. Similarly we define respectively $\ddot{m}_\theta$ and $\dddot{m}_\theta$ as the second and third partial derivatives of m_θ with respect to θ. For any vector z we define $z^{\otimes 2} \triangleq zz^\top$.

4.3.1 Intuition for Square-Root Importance Weights

The critical role of the square-root importance weights $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t,X_t)}{\pi_t(A_t,X_t,\mathcal{H}_{t-1})}}$ is to adjust for instability in the *variance* of M-estimators due to the bandit algorithm. These weights act akin to standard importance weights when squared and adjust a key term in the variance of M-estimators from depending on adaptive policies $\{\pi_t\}_{t=1}^T$, which can be ill-behaved, to depending on the pre-specified stabilizing policies $\{\pi_t^{\text{sta}}\}_{t=1}^T$. See Zhang et al. [109] and Deshpande et al. [26] for more discussion of

the ill-behavior of the action selection probabilities for common bandit algorithms, which occurs particularly when there is no unique optimal policy.

As an illustrative example, consider the least-squares estimators in a finite-arm linear contextual bandit setting. Assume that $\mathbb{E}_{\mathcal{P}}[Y_t|X_t, A_t = a] = X_t^\top \theta_a^*(\mathcal{P})$ w.p. 1. We focus on estimating $\theta_a^*(\mathcal{P})$ for some $a \in \mathcal{A}$. The least-squares estimator corresponds to an M-estimator with $m_{\theta_a}(Y_t, X_t, A_t) = -1_{A_t=a}(Y_t - X_t^\top \theta_a)^2$. The adaptively weighted least-squares (AW-LS) estimator is $\hat{\theta}_{T,a}^{\text{AW-LS}} \triangleq \text{argmax}_{\theta_a}\{-\sum_{t=1}^T W_t 1_{A_t=a}(Y_t - X_t^\top \theta_a)^2\}$. For simplicity, suppose that the stabilizing policy does not change with t and drop the index t to get π^{sta}. Taking the derivative of this criterion, we get $0 = \sum_{t=1}^T W_t 1_{A_t=a} X_t \big(Y_t - X_t^\top \hat{\theta}_{T,a}^{\text{AW-LS}}\big)$, and rearranging terms gives

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T W_t 1_{A_t=a} X_t X_t^\top \left(\hat{\theta}_{T,a}^{\text{AW-LS}} - \theta_a^*(\mathcal{P})\right) = \frac{1}{\sqrt{T}} \sum_{t=1}^T W_t 1_{A_t=a} X_t \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right). \quad (4.3.3)$$

Note that the right hand side of Equation (4.3.3) is a martingale difference sequence with respect to history $\{\mathcal{H}_t\}_{t=0}^T$ because $\mathbb{E}_{\mathcal{P},\pi}[W_t 1_{A_t=a}(Y_t - X_t^\top \theta_a^*(\mathcal{P}))|\mathcal{H}_{t-1}] = 0$ for all t; by law of iterated expectations and since $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$, $\mathbb{E}_{\mathcal{P},\pi}[W_t 1_{A_t=a}(Y_t - X_t^\top \theta_a^*(\mathcal{P}))|\mathcal{H}_{t-1}]$ equals

$$\mathbb{E}_{\mathcal{P}}\left[W_t \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}}\left[Y_t - X_t^\top \theta_a^*(\mathcal{P})|\mathcal{H}_{t-1}, X_t, A_t = a\right] \big|\mathcal{H}_{t-1}\right]$$

$$\underset{(i)}{=} \mathbb{E}_{\mathcal{P}}\left[W_t \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}}\left[Y_t - X_t^\top \theta_a^*(\mathcal{P})|X_t, A_t = a\right] \big|\mathcal{H}_{t-1}\right] \underset{(ii)}{=} 0.$$

(i) holds by our i.i.d. potential outcomes assumption. (ii) holds since $\mathbb{E}_{\mathcal{P}}[Y_t|X_t, A_t = a] = X_t^\top \theta_a^*(\mathcal{P})$. We prove that (4.3.3) is uniformly asymptotically normal by applying a martingale central limit theorem (Appendix B.2.4). The key condition in this theorem is that the conditional variance converges uniformly, for which it is sufficient to show that the conditional covariance of

$W_t 1_{A_t=a}\left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)$ given $\mathcal{H}_{t-1}$ equals some positive-definite matrix $\Sigma(\mathcal{P})$ for every t, i.e.,

$$\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 1_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \middle| \mathcal{H}_{t-1}\right] = \Sigma(\mathcal{P}). \tag{4.3.4}$$

By law of iterated expectations, $\mathbb{E}_{\mathcal{P},\pi}[W_t^2 1_{A_t=a} X_t X_t^\top (Y_t - X_t^\top \theta_a^*(\mathcal{P}))^2 \big| \mathcal{H}_{t-1}]$ equals

$$\mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi}\left[\frac{\pi^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})} 1_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \middle| \mathcal{H}_{t-1}, X_t\right] \middle| \mathcal{H}_{t-1}\right] \tag{4.3.5}$$

$$\underset{(a)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}\left[1_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \middle| \mathcal{H}_{t-1}, X_t\right] \middle| \mathcal{H}_{t-1}\right]$$

$$\underset{(b)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}\left[1_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \middle| X_t\right] \middle| \mathcal{H}_{t-1}\right]$$

$$\underset{(c)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}\left[1_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \middle| X_t\right]\right]$$

$$\underset{(d)}{=} \mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}[1_{A_t=a} X_t X_t^\top (Y_t - X_t^\top \theta_a^*(\mathcal{P}))^2] =: \Sigma(\mathcal{P}).$$

Above, (a) holds because the importance weights change the sampling measure from the adaptive policy π_t to the pre-specified stabilizing policy π^{sta}. (b) holds by our i.i.d. potential outcomes assumption and because π^{sta} is a pre-specified policy. (c) holds because X_t does not depend on $\mathcal{H}_{t-1}$ by our i.i.d. potential outcomes assumption. (d) holds by the law of iterated expectations. Note that $\Sigma(\mathcal{P})$ does not depend on t because π^{sta} is not time-varying. In contrast, without the adaptive weighting, i.e., when $W_t = 1$, the conditional covariance of $1_{A_t=a}\left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)$ on $\mathcal{H}_{t-1}$ is a random variable, due to the adaptive policy π_t.

In Figure 4.1 we plot the empirical distributions of the z-statistic for the least-squares estimator both with and without adaptive weighting. We consider a two-armed bandit with $A_t \in \{0, 1\}$. Let $\theta_1^*(\mathcal{P}) \triangleq \mathbb{E}_{\mathcal{P}}[Y_t(1)]$ and $m_{\theta_1}(Y_t, A_t) \triangleq -A_t(Y_t - \theta_1)^2$. The unweighted version, i.e., the ordinary least-squares (OLS) estimator, is $\hat{\theta}_{T,1}^{\text{OLS}} \triangleq \text{argmax}_{\theta_1} \frac{1}{T}\sum_{t=1}^T m_{\theta_1}(Y_t, A_t)$. The adaptively weighted

version is $\hat{\theta}^{\text{AW-LS}}_{T,1} \triangleq \text{argmax}_{\theta_1} \frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta_1}(Y_t, A_t)$. We collect data using Thompson Sampling and use a uniform stabilizing policy where $\pi^{\text{sta}}(1) = \pi^{\text{sta}}(0) = 0.5$. It is clear that the least-squares estimator with adaptive weighting has a z-statistic that is much closer to a normal distribution.

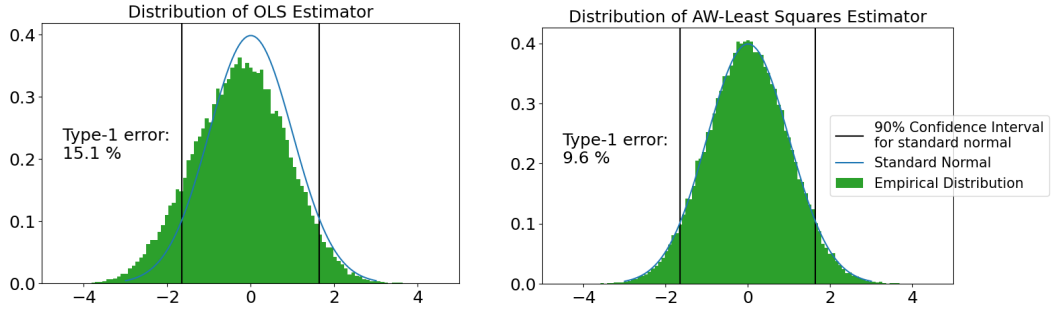

Figure 4.1: The empirical distributions of the weighted and unweighted least-squares estimators for $\theta_1^*(\mathcal{P}) \triangleq \mathbb{E}_{\mathcal{P}}[Y_t(1)]$ in a two arm bandit setting where $\mathbb{E}_{\mathcal{P}}[Y_t(1)] = \mathbb{E}_{\mathcal{P}}[Y_t(0)] = 0$. We use Thompson Sampling with $\mathcal{N}(0,1)$ priors, $\mathcal{N}(0,1)$ errors, and $T = 1000$. We plot $\sqrt{\sum_{t=1}^{T} A_t}(\hat{\theta}^{\text{OLS}}_{T,1} - \theta_1^*(\mathcal{P}))$ on the left and $\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sqrt{\frac{0.5}{\pi_t(1)}}A_t\right)(\hat{\theta}^{\text{AW-LS}}_{T,1} - \theta_1^*(\mathcal{P}))$ on the right.

The square-root importance weights are a form of variance stabilizing weights, akin to those introduced in Hadad et al. [39] for estimating means and differences in means on data collected with multi-armed bandits. In fact, in the special case in which the action space is finite ($|\mathcal{A}| < \infty$) and $\varphi(X_t, A_t) = [1_{A_t=1}, 1_{A_t=2}, \ldots, 1_{A_t=|\mathcal{A}|}]^\top$, the adaptively weighted least-squares estimator is equivalent to the weighted average estimator of Hadad et al. [39]. See Section 4.4 for more on Hadad et al. [39].

4.3.2 Asymptotic Normality and Confidence Regions

We now discuss conditions under which the adaptively weighted M-estimators are asymptotically normal in the sense of Equation (4.3.2). In general, our conditions differ from those made for standard M-estimators on i.i.d. data because (i) the data is adaptively collected, i.e., π_t can depend on $\mathcal{H}_{t-1}$ and (ii) we ensure uniform convergence over $\mathcal{P} \in \mathbf{P}$, which is stronger than guaranteeing convergence pointwise for each $\mathcal{P} \in \mathbf{P}$.

Condition 4.3.1 (Stochastic Bandit Environment)**.** *Potential outcomes* $\{X_t, Y_t(a) : a \in \mathcal{A}\} \overset{i.i.d.}{\sim} \mathcal{P} \in \boldsymbol{P}$ *over* $t \in [1\colon T]$.

Condition 4.3.1 implies that Y_t is independent of $\mathcal{H}_{t-1}$ given X_t and A_t, and the conditional distribution $Y_t \mid X_t, A_t$ is invariant over time. Also note that action space $\mathcal{A}$ can be finite or infinite.

Condition 4.3.2 (Differentiable)**.** *The first three derivatives of* $m_\theta(y, x, a)$ *with respect to* θ *exist for every* $\theta \in \Theta$*, every* $a \in \mathcal{A}$*, and every* (x, y) *in the joint support of* $\{\mathcal{P} : \mathcal{P} \in \boldsymbol{P}\}$.

Condition 4.3.3 (Bounded Parameter Space)**.** *For all* $\mathcal{P} \in \boldsymbol{P}$*,* $\theta^*(\mathcal{P}) \in \Theta$*, a bounded open subset of* $\mathbb{R}^d$.

Condition 4.3.4 (Lipschitz)**.** *There exists some real-valued function g such that*

(i) $\sup_{\mathcal{P}\in\boldsymbol{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[g(Y_t, X_t, A_t)^2]$ *is bounded and (ii) for all* $\theta, \theta' \in \Theta$*,*

$$|m_\theta(Y_t, X_t, A_t) - m_{\theta'}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t)\|\theta - \theta'\|_2.$$

Conditions 4.3.3 and 4.3.4 together restrict the complexity of the function m in order to ensure a martingale law of large numbers result holds uniformly over functions $\{m_\theta : \theta \in \Theta\}$; this is used to prove the consistency of $\hat{\theta}_T$. Similar conditions are commonly used to prove consistency of M-estimators based on i.i.d. data, although the boundedness of the parameter space can be dropped when m_θ is a concave function of θ for all Y_t, A_t, X_t (as it is in many canonical examples such as least squares)[99,31,18]; we expect that a similar result would hold for adaptively weighted M-estimators.

Condition 4.3.5 (Moments)**.** *The fourth moments of* $m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)$*,* $\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)$*, and* $\ddot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)$ *with respect to* $\mathcal{P}$ *and policy* π_t^{sta} *are bounded uniformly over* $\mathcal{P} \in \boldsymbol{P}$ *and* $t \geq 1$*. The minimum eigenvalue of* $\Sigma_{T,P} \triangleq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)^{\otimes 2}\right]$ *is bounded above* $\delta_{\dot{m}^2} > 0$ *for all* $\mathcal{P} \in \boldsymbol{P}$ *and all sufficiently large T.*

Condition 4.3.5 is similar to those of Van der Vaart [99], Theorem 5.41. However, to guarantee uniform convergence we assume that moment bounds hold uniformly over $\mathcal{P} \in \mathbf{P}$ and $t \geq 1$.

Condition 4.3.6 (Third Derivative Domination)**.** *For any $B \in \mathbb{R}^{d\times d\times d}$, we define $\|B\|_1 \triangleq \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d |B_{i,j,k}|$. There exists a function $\dddot{m}(Y_t, X_t, A_t) \in \mathbb{R}^{d\times d\times d}$ such that*

(i) $\sup_{\mathcal{P}\in\boldsymbol{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\|\dddot{m}(Y_t, X_t, A_t)\|_1^2\right]$ *is bounded and*

(ii) for all $\mathcal{P} \in \boldsymbol{P}$ there exists some $\varepsilon_{\dddot{m}} > 0$ such that the following holds with probability 1*,*

$$\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|\leq\varepsilon_{\dddot{m}}} \|\dddot{m}_\theta(Y_t, X_t, A_t)\|_1 \leq \|\dddot{m}(Y_t, X_t, A_t)\|_1.$$

Condition 4.3.6 is again similar to those in classical M-estimator asymptotic normality proofs [99] Theorem 5.41.

Condition 4.3.7 (Maximizing Solution)**.**

(i) For all $\mathcal{P} \in \boldsymbol{P}$, there exists a $\theta^*(\mathcal{P}) \in \Theta$ such that

(a) $\theta^*(\mathcal{P}) \in \operatorname{argmax}_{\theta\in\Theta} \mathbb{E}_{\mathcal{P}}\left[m_\theta(Y_t, X_t, A_t)\middle|X_t, A_t\right]$ *w.p.* 1*,*

(b) $\mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\middle|X_t, A_t\right] = 0$ *w.p.* 1*, and*

(c) $\mathbb{E}_{\mathcal{P}}\left[\ddot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\middle|X_t, A_t\right] \preceq 0$ *w.p.* 1*.*

(ii) There exists some positive definite matrix H such that $-\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\ddot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\right] \succeq H$ for all $\mathcal{P} \in \boldsymbol{P}$ and all sufficiently large T.

For matrices A, B, we define $A \succeq B$ to mean that $A - B$ is positive semi-definite, as used above. Condition 4.3.7 (i) ensures that $\theta^*(\mathcal{P})$ is a conditionally maximizing solution for all contexts X_t and actions A_t; this ensures that $\{\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t=1}^T$. Note it does not require $\theta^*(\mathcal{P})$ to always be a conditionally *unique* optimal solution. Condition 4.3.7 (ii) is related to the local curvature at the maximizing solution and the analogous condition in the i.i.d. setting is trivially satisfied; we specifically use this condition to ensure we can

replace $\ddot{M}(\theta^*(\mathcal{P}))$ with $\ddot{M}(\hat{\theta}_T)$ in our asymptotic normality result, i.e., that $\ddot{M}(\theta^*(\mathcal{P}))^{-1}\ddot{M}(\hat{\theta}_T) \xrightarrow{P} I_d$ uniformly over $\mathcal{P} \in \mathbf{P}$.

Condition 4.3.8 (Well-Separated Solution). *For all sufficiently large T, for any $\varepsilon > 0$, there exists some $\delta > 0$ such that for all $\mathcal{P} \in \boldsymbol{P}$,*

$$\inf_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\|_2 > \varepsilon} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t) - m_\theta(Y_t, X_t, A_t) \right] \right\} \geq \delta.$$

A well-separated solution condition akin to Condition 4.3.8 is commonly assumed in order to prove consistency of M-estimators, e.g., see Van der Vaart [99], Theorem 5.7. Note that the difference between Condition 4.3.7 (i) and Condition 4.3.8 is that the former is a conditional statement (conditional on X_t, A_t) and the latter is a marginal statement (marginal over X_t, A_t, where A_t is chosen according to stabilizing policies π_t^{sta}). Condition 4.3.7 (i) means there is a $\theta^*(\mathcal{P})$ solution for all contexts X_t and actions A_t that does not need to be unique, however Condition 4.3.8 assumes that marginally over X_t, A_t there is a well-separated solution.

Condition 4.3.9 (Bounded Importance Ratios). *$\{\pi_t^{\text{sta}}\}_{t=1}^T$ do not depend on data $\{Y_t, X_t, A_t\}_{t=1}^T$. For all $t \geq 1, \rho_{\min} \leq \frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})} \leq \rho_{\max}$ w.p. 1 for some constants $0 < \rho_{\min} \leq \rho_{\max} < \infty$.*

Note that Condition 4.3.9 implies that for a stabilizing policy that is not time-varying, the action selection probabilities of the bandit algorithm $\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ must be bounded away from zero w.p. 1. Similar boundedness assumptions are also made in the off-policy evaluation literature [94,53]. We discuss this condition further in Sections 4.3.3 and 4.6.

Theorem 4.3.1 (Uniform Asymptotic Normality of Adaptively Weighted M-Estimators). *Under Conditions 4.3.1-4.3.9 we have that $\hat{\theta}_T \xrightarrow{P} \theta^*(\mathcal{P})$ uniformly over $\mathcal{P} \in \boldsymbol{P}$. Additionally,*

$$\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\hat{\theta}_T) \sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \boldsymbol{P}. \tag{4.3.6}$$

The asymptotic normality result of equation (4.3.6) guarantees that for d-dimensional $\theta^*(\mathcal{P})$,

$$\liminf_{T\to\infty} \inf_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\left[\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P}))\right]^{\otimes 2} \leq \chi^2_{d,(1-\alpha)}\right) = 1-\alpha.$$

Above $\chi^2_{d,(1-\alpha)}$ is the $1-\alpha$ quantile of the χ^2 distribution with d degrees of freedom. Note that the region $C_T(\alpha) \triangleq \left\{\theta \in \Theta : [\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P}))]^{\otimes 2} \leq \chi^2_{d,(1-\alpha)}\right\}$ defines a d-dimensional hyper-ellipsoid confidence region for $\theta^*(\mathcal{P})$. Also note that since $\ddot{M}_T(\hat{\theta}_T)$ does not concentrate under standard bandit algorithms, we cannot use standard arguments to justify treating $\hat{\theta}_T$ as multivariate normal with covariance $\ddot{M}_T(\hat{\theta}_T)^{-1}\Sigma_T(\mathcal{P})\ddot{M}_T(\hat{\theta}_T)^{-1}$. Nevertheless, Theorem 4.3.1 can be used to guarantee valid confidence regions for subset of entries in $\theta^*(\mathcal{P})$ by using projected confidence regions[75]. Projected confidence regions take a confidence region for all parameters $\theta^*(\mathcal{P})$ and project it onto the lower dimensional space on which the subset of target parameters lie (Appendix B.1.2).

4.3.3 Choice of Stabilizing Policy

When the action space is bounded, using weights $W_t = 1/\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$ is equivalent to using square-root importance weights with a stabilizing policy that selects actions uniformly over $\mathcal{A}$; this is because weighted M-estimators are invariant to all weights being scaled by the same constant. It can make sense to choose a non-uniform stabilizing policy in order to prevent the square-root importance weights from growing too large and to ensure Condition 4.3.9 holds; disproportionately up-weighting a few observations can lead to unstable estimators. Note that an analogue of our stabilizing policy exists in the causal inference literature, namely, "stabilized weights" use a probability density in the numerator of the weights to prevent them from becoming too large[85].

We now discuss how to choose stabilizing policies $\{\pi_t^{\text{sta}}\}_{t\geq 1}$ in order to minimize the asymptotic variance of adaptively weighted M-estimators. We focus on the adaptively weighted least-squares

estimator when we have a linear outcome model $\mathbb{E}_{\mathcal{P}}[Y_t|X_t, A_t] = X_t^\top \theta_{A_t}$:

$$\hat{\theta}^{\text{AW-LS}} \triangleq \underset{\theta\in\Theta}{\operatorname{argmax}} \left\{ \frac{1}{T} \sum_{t=1}^{T} W_t \left(Y_t - X_t^\top \theta_{A_t} \right)^2 \right\}. \tag{4.3.7}$$

Recall that our use of adaptive weights is to adjust for instability in the variance of M-estimators induced by the bandit algorithm in order to construct valid confidence regions; note that weighted estimators are not typically used for this reason. On i.i.d. data, the least-squares criterion is weighted like in Equation (4.3.7) in order to minimize the variance of estimators under noise heteroskedasticity; in this setting, the best linear unbiased estimator has weights $W_t = 1/\sigma^2(A_t, X_t)$ where $\sigma^2(A_t, X_t) \triangleq \mathbb{E}_{\mathcal{P}}[(Y_t - X_t^\top \theta^*_{A_t}(\mathcal{P}))^2 | X_t, A_t]$; this up-weights the importance of observations with low noise variance. Intuitively, if we do not need to variance stabilize, $\{W_t\}_{t\geq 1}$ should be determined by the relative importance of minimizing the errors for different observations, i.e., their noise variance.

In light of this observation, we expect that under homoskedastic noise there is no reason to up-weight some observations over others. This would recommend choosing the stabilizing policy to make $W_t = \sqrt{\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$ as close to 1 as possible, subject to the constraint that the stabilizing policies are pre-specified, i.e., $\{\pi_t^{\text{sta}}\}_{t\geq 1}$ do not depend on data $\{Y_t, X_t, A_t\}_{t\geq 1}$ (see Appendix B.3 for details). Since adjusting for heteroskedasticity and variance stabilization are distinct uses of weights, under heteroskedasticity, we recommend that the weights are combined in the following sense: $W_t = \left(1/\sigma^2(A_t, X_t)\right) \sqrt{\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$. This would mean that to minimize variance, we still want to choose the stabilizing policies in order to make $\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ as close to 1 possible, subject to the pre-specified constraint.

4.4 Related Work

Villar et al. [101] and Rafferty et al. [82] empirically illustrate that classical ordinary least squares (OLS)

inference methods have inflated Type-1 error when used on data collected with a variety of regret-minimizing multi-armed bandit algorithms. Chen et al. [23] prove that the OLS estimator is asymptotically normal on data collected with an ε-greedy algorithm, but their results do not cover settings in which there is no unique optimal policy, e.g., a multi-arm bandit with two identical arms (Appendix B.5). Recent work has discussed the non-normality of OLS on data collected with bandit algorithms when there is no unique optimal policy and proposed alternative methods for statistical inference. A common thread between these methods is that they all utilize a form of *adaptive weighting*. Deshpande et al. [26] introduced the W-decorrelated estimator, which adjusts the OLS estimator with a sum of adaptively weighted residuals. In the multi-armed bandit setting, the W-decorrelated estimator up-weights observations from early in the study and down-weights observations from later in the study [109]. In the batched bandit setting, Zhang et al. [109] show that the Z-statistics for the OLS estimators computed separately on each batch are jointly asymptotically normal. Standardizing the OLS statistic for each batch effectively adaptively re-weights the observations in each batch.

Hadad et al. [39] introduce adaptively weighted versions of both the standard augmented-inverse propensity weighted estimator (AW-AIPW) and the sample mean (AWA) for estimating parameters of simple models on data collected with bandit algorithms. They introduce a class of adaptive "variance stabilizing" weights, for which the variance of a normalized version of their estimators converges in probability to a constant. In their discussion section they note open questions, two of which this work addresses: 1) "What additional estimators can be used for normal inference with adaptively collected data?" and 2) How do their results generalize to more complex sampling designs, like data collected with contextual bandit algorithms? We demonstrate that variance stabilizing adaptive weights can be used to modify a large class of M-estimators to guarantee valid inference. This generalization allows us to perform valid inference for a large class of important inferential targets: parameters of models for expected outcomes that are context dependent.

Recently, adaptive weighting has also been used in off-policy evaluation methods for when the

behavior policy (policy used to collect the data) is a contextual bandit algorithm [11,108]. In this literature the estimand is the value, or average expected reward, of a pre-specified policy (note this is a scalar value). In contrast, in our work we are interested in constructing confidence regions for parameters of a model for an outcome (that could be the reward)—for example, this could be parameters of a logistic regression model for a binary outcome. We believe in the future there could be theory that could unify these adaptive weighting methods for these different estimands.

An alternative to using asymptotic approximations to construct confidence intervals is to use high-probability confidence bounds. These bounds provide stronger guarantees than those based on asymptotic approximations, as they are guaranteed to hold for finite samples. The downside is that these bounds are typically much wider, which is why much of classical statistics uses asymptotic approximations. Here we do the same. In Section 4.5, we empirically compare our to the self-normalized martingale bound [2], a high-probability bound commonly used in the bandit literature.

4.5 Simulation Results

In this section, $R_t = Y_t$. We consider two settings: a continuous reward setting and a binary reward setting. In the continuous reward setting, the rewards are generated with mean $\mathbb{E}_{\mathcal{P}}[R_t|X_t, A_t] = \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$ and noise drawn from a student's t distribution with five degrees of freedom; here $\tilde{X}_t = [1, X_t] \in \mathbb{R}^3$ (X_t with intercept term), actions $A_t \in \{0, 1\}$, and parameters $\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P}) \in \mathbb{R}^3$. In the binary reward setting, the reward R_t is generated as a Bernoulli with success probability $\mathbb{E}_{\mathcal{P}}[R_t|X_t, A_t] = [1 + \exp(-\tilde{X}_t^\top \theta_0^*(\mathcal{P}) - A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P}))]^{-1}$. Furthermore, in both simulation settings we set $\theta_0^*(\mathcal{P}) = [0.1, 0.1, 0.1]$ and $\theta_1^*(\mathcal{P}) = [0, 0, 0]$, so there is no unique optimal arm; we call vector parameter $\theta_1^*(\mathcal{P})$ the *advantage* of selecting $A_t = 1$ over $A_t = 0$. Also in both settings, the contexts X_t are drawn i.i.d. from a uniform distribution.

In both simulation settings we collect data using Thompson Sampling with a linear model for

the expected reward and normal priors[4] (so even when the reward is binary). We constrain the action selection probabilities with *clipping* at a rate of 0.05; this means that while typical Thompson Sampling produces action selection probabilities $\pi_t^{\mathrm{TS}}(A_t, X_t, \mathcal{H}_{t-1})$, we instead use action selection probabilities $\pi_t(A_t, X_t, \mathcal{H}_{t-1}) = 0.05 \vee \left(0.95 \wedge \pi_t^{\mathrm{TS}}(A_t, X_t, \mathcal{H}_{t-1})\right)$ to select actions. We constrain the action selection probabilities in order to ensure weights W_t are bounded when using a uniform stabilizing policy; see Sections 4.3.2 and 4.6 for more discussion on this boundedness assumption. Also note that increasing the amount the algorithm explores (clipping) decreases the expected width of confidence intervals constructed on the resulting data (see Section 4.6).

To analyze the data, in the continuous reward setting, we use least-squares estimators with a correctly specified model for the expected reward, i.e., M-estimators with $m_\theta(R_t, X_t, A_t) = -(R_t - \tilde{X}_t^\top \theta_0 - A_t \tilde{X}_t^\top \theta_1)^2$. We consider both the unweighted and adaptively weighted versions. We also compare to the self-normalized martingale bound[2] and the W-decorrelated estimator[26], as they were both developed for the linear expected reward setting. For the self-normalized martingale bound, which requires explicit bounds on the parameter space, we set $\Theta = \{\theta \in \mathbb{R}^6 : \|\theta\|_2 \leq 6\}$. In the binary reward setting, we also assume a correctly specified model for the expected reward. We use both unweighted and adaptively weighted maximum likelihood estimators (MLEs), which correspond to an M-estimators with $m_\theta(R_t, X_t, A_t)$ set to the negative log-likelihood of R_t given X_t, A_t. We solve for these estimators using Newton–Raphson optimization and do not put explicit bounds on the parameter space Θ (note in this case m_θ is concave in θ[5] Chapter 5.4.2). See Appendix B.1 for additional details and simulation results.

In Figure 4.2 we plot the empirical coverage probabilities and volumes of 90% confidence regions for $\theta^*(\mathcal{P}) \triangleq [\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})]$ and $\theta_1^*(\mathcal{P})$ in both the continuous and binary reward settings. While the confidence regions based on the unweighted least-squares estimator (OLS) and the unweighted MLE have significant undercoverage that does not improve as T increases, the confidence regions based on the adaptively weighted versions, AW-LS and AW-MLE, have very reliable coverage. For the

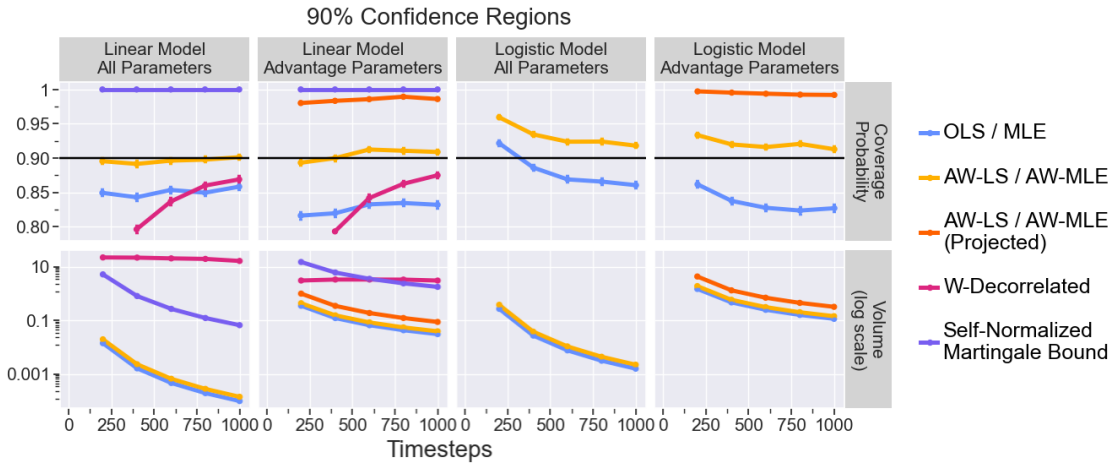

Figure 4.2: Empirical coverage probabilities (upper row) and volume (lower row) of 90% confidence ellipsoids. We consider both the linear reward model setting with t-distributed rewards (left two columns) and the logistic regression model setting with binary rewards (right two columns). We consider confidence ellipsoids for all parameters $\theta^*(\mathcal{P})$ and for advantage parameters $\theta_1^*(\mathcal{P})$ for both settings.

confidence regions for $\theta_1^*(\mathcal{P})$ based on the AW-LS and AW-MLE, we include both projected confidence regions (for which we have theoretical guarantees) and non-projected confidence regions. The confidence regions based on projections are conservative but nevertheless have comparable volume to those based on OLS and MLE respectively. We do not prove theoretical guarantees for the non-projection confidence regions for AW-LS and AW-MLE, however they perform well across in our simulations. Both types of confidence regions based on AW-LS have significantly smaller volumes than those constructed using the self-normalized martingale bound and W-decorrelated estimator. Note that the W-decorrelated estimator and self-normalized martingale bounds are designed for linear contextual bandits and are thus not applicable for the logistic regression model setting. The confidence regions constructed using the self-normalized martingale bound have reliable coverage as well, but are very conservative. Empirically, we found that the coverage probabilities of the confidence regions based on the W-decorrelated estimator were very sensitive to the choice of tuning parameters. We use $5,000$ Monte-Carlo repetitions and the error bars plotted are standard errors.

4.6 Discussion

Immediate questions We assume that ratios $\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ are bounded for our theoretical results; this precludes $\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ from going to zero for a fixed stabilizing policy. For simple models, e.g., the AW-LS estimator, we can let these ratios grow at a certain rate and still guarantee asymptotic normality (Appendix B.2.5); we conjecture similar results hold more generally.

Generality and robustness This work assumes that we have a well-specified model for the outcome Y_t, i.e., that $\theta^*(\mathcal{P}) \in \operatorname{argmax}_{\theta\in\Theta} \mathbb{E}_{\mathcal{P}}[m_\theta(Y_t, X_t, A_t)|X_t, A_t]$ w.p. 1. Our theorems use this assumption to ensure that $\{W_t \dot{m}_\theta(Y_t, X_t, A_t)\}_{t\geq 1}$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t\geq 0}$. On i.i.d. data it is common to define $\theta^*(\mathcal{P})$ to be the best *projected* solution, i.e., $\theta_0(\mathcal{P}) \in \operatorname{argmax}_{\theta\in\Theta} \mathbb{E}_{\mathcal{P},\pi}[m_\theta(Y_t, X_t, A_t)]$. Note that the best projected solution, $\theta^*(\mathcal{P})$, depends on the distribution of the action selection policy π. It would be ideal to also be able to perform inference for a projected solution on adaptively collected data.

Another natural question is whether adaptive weighting methods work in Markov Decision Processes (MDP) environments. Taking the AW-LS estimator introduced in Section 4.3.1 as an example, our conditional variance derivation in Equation (4.3.5) fails to hold in an MDP setting, specifically equality (c). However, the conditional variance condition can be satisfied if we instead use weights $W_t = \{[\pi_t^{\text{sta}}(A_t, X_t)p^{\text{sta}}(X_t)]/[\pi_t(A_t, X_t, \mathcal{H}_{t-1})\mathbb{P}_{\mathcal{P}}(X_t|X_{t-1}, A_{t-1})]\}^{1/2}$ where $\mathbb{P}_{\mathcal{P}}$ are the state transition probabilities and p^{sta} is a pre-specified distribution over states. In general though we do not expect to know the transition probabilities $\mathbb{P}_{\mathcal{P}}$ and if we tried to estimate them, our theory would require the estimator to have error $o_p(1/\sqrt{T})$, *below* the parametric rate.

Trading-off regret minimization and statistical inference objectives In sequential decision-making problems there is a fundamental trade-off between minimizing regret and minimizing estimation error for parameters of the environment using the resulting data [17,24]. Given this trade-off there are many open problems regarding how to minimize regret while still guaranteeing a certain

amount of power or expected confidence interval width, e.g., developing sample size calculators for use in justifying the number of users in a mobile health trial, and developing new adaptive algorithms[70,32,104].

5

Inference after Adaptive Sampling for Longitudinal Data

Online reinforcement learning and other adaptive sampling algorithms are increasingly used in digital intervention experiments to optimize treatment delivery for users over time. In this work, we focus on longitudinal user data collected by a large class of adaptive sam-

pling algorithms that are designed to optimize treatment decisions online using accruing data from multiple users. Combining or "pooling" data across users allows adaptive sampling algorithms to potentially learn faster. However, by pooling, these algorithms induce dependence between the sampled user data trajectories; we show that this can cause standard variance estimators for i.i.d. data to underestimate the true variance of common estimators on this data type. We develop novel methods to perform a variety of statistical analyses on such adaptively sampled data via Z-estimation. Specifically, we introduce the *adaptive* sandwich variance estimator, a corrected sandwich estimator that leads to consistent variance estimates under adaptive sampling. Additionally, to prove our results we develop novel theoretical tools for empirical processes on non-i.i.d., adaptively sampled longitudinal data which may be of independent interest. This work is motivated by our efforts in designing experiments in which online reinforcement learning algorithms optimize treatment decisions, yet statistical inference is essential for conducting analyses after experiments conclude.

5.1 Introduction

Online reinforcement learning (RL) and other adaptive sampling algorithms are increasingly used in digital intervention experiments to optimize treatment delivery for users over time [34,36,67,79,96,97,105]. For example, in mobile health, online RL algorithms have been used in longitudinal clinical trials developing Just-In-Time interventions for people with a variety of chronic health problems [34,67,79,105]. These trials are longitudinal because they involve making multiple treatment decisions for users over time. Online RL algorithms are used during the experiment to optimize treatment decisions; specifically, the RL algorithm uses the outcomes of previous treatment decisions to inform future treatment selection. A critical consideration when designing experiments that use such RL algorithms is ensuring that one can use the resulting data collected to perform valid statistical inference after the experiment is over. For example, one might want to construct confidence intervals for a

treatment effect on a variety of different outcomes, including the reward. This work is motivated by experimental trials for digital health interventions in which RL algorithms are used to optimize treatments over time for multiple users. In these settings, valid post-trial statistical inference is critical to inform decisions about whether to roll out or how to improve a given digital intervention after the trial is over[34,67,96]. A significant challenge is to develop valid statistical inference methods that are applicable to the data collected by the variety of RL algorithms that those designing digital interventions want to use.

Recently in the longitudinal digital intervention space, there has been great interest in online RL algorithms that combine or "pool" data across multiple users to inform future treatment decisions, because they can potentially learn faster how best to select treatments. In fact, there is so much interest that several digital health intervention trials have already used such pooling RL algorithms[34,79,95,105]. However, it is unclear whether existing statistical inference methods for longitudinal data[13,35,81,84,106], which assume independent user data trajectories, should be used on data collected with adaptive sampling algorithms that pool data online. This is because, by pooling, adaptive sampling algorithms induce dependence between the collected user data trajectories. For example, if the algorithm uses the outcomes of one user to inform future treatment decisions for another user, the data trajectories collected from these two users will not be independent.

There are existing approaches for statistical inference after adaptive sampling that account for the dependence induced by the algorithm. However, these approaches make a variety of restrictive assumptions on how users' outcomes can evolve over time and can be affected by treatments. For example, many works assume a classical contextual bandit environment in which user states are i.i.d. over time and the mean reward only depends on the most recent state and treatment[11,39,44,108,109,110]. These inference approaches are not applicable to classical longitudinal data settings in which (a) treatment decisions may affect users' future responsiveness to treatments or (b) user outcomes may be non-stationary.

Moreover, we want our statistical inference approach to be robust to misspecification of the model used by the RL algorithm. Specifically, online RL algorithms make treatment decisions using approximate models for the users' outcomes (i.e., models of the environment) that they repeatedly fit using the data collected during the experiment. The models used in online RL algorithms are chosen to appropriately trade off bias and variance so the algorithm can quickly learn to select effective treatments. For example, even in environments in which treatment decisions may impact users' responsiveness to treatments multiple decision times into the future, in order to reduce variance, often simpler algorithms that do not model these delayed effects of treatment (like bandit algorithms) are preferred[34,96,105]. After the digital intervention experiment is over, we argue that the validity of the statistical inference using the resulting adaptively sampled longitudinal data should not require that these approximate models used by the online RL algorithm are correctly specified.

5.1.1 Our Contribution

In this work, we consider pooling adaptive sampling algorithms that, for each decision time $t \in [1:T]$, form a policy $\hat{\pi}_t^{(n)}$ that appropriately converges to a *target* policy π_t^* as the number of users n grows; see Remark 5.2.2 and Section 5.3.2 for further discussion of this assumption. We provide statistical theory for Z-estimators[99] Chapter 5 on data collected by such pooling adaptive sampling algorithms. Z-estimators encompass most classical statistical estimators (e.g., least squares and maximum likelihood estimators) and are often used in estimating time-varying causal effects[84]. We derive the asymptotic distribution of these Z-estimators as the number of users $n \to \infty$ to facilitate the construction of asymptotically valid confidence regions. Specifically, we prove that the commonly used standard sandwich variance estimator[45,107], can underestimate the true variance of Z-estimators when data is adaptively sampled via algorithms that learn by pooling data across users. We develop the *adaptive sandwich estimator*, a corrected sandwich estimator that leads to consistent variance estimates under adaptive sampling. Specifically, our contributions are as follows:

1. **Facilitating Statistical Inference after Using Pooling Adaptive Sampling Algorithms on Longitudinal Data:** Our approach for inference via Z-estimators is the first method that is applicable to longitudinal datasets collected by adaptive sampling algorithms that learn by pooling data across users. Moreover, the validity of our approach does not require the approximate outcome models learned by the adaptive sampling algorithm to be correct. This work enables digital intervention researchers to use pooling adaptive sampling algorithms in their experiments without sacrificing the statistical validity in performing a wide variety of after-study analyses.

2. **Novel use of Radon-Nikodym Derivatives to Facilitate Theory for Adaptively Sampled Data:** A significant technical challenge is that standard methods for empirical processes are insufficient for proving our asymptotic normality results since the adaptively sampled user data trajectories are not i.i.d. A key approach we use to facilitate theory for this non-i.i.d. data type is Radon-Nikodym derivative weighting. Specifically, we consider settings in which the estimator of the parameter of interest and the estimators used by the adaptive sampling policies $\hat{\pi}_t^{(n)}$ are each a solution to some standard estimating function. Incorporating Radon-Nikodym derivative weights is integral to defining *joint* estimating functions for the parameter of interest and the policy parameters. Note that the joint estimating functions (and the Radon-Nikodym derivative weights) are used *solely* to analyze the asymptotic distribution of these estimators and *not* needed to form the estimators themselves. We introduce these weights in Section 5.5.1.

3. **Empirical Process Theory for Adaptively Sampled Longitudinal Data:** To prove our results we develop novel theoretical tools for empirical processes on non-i.i.d., adaptively sampled longitudinal data, which may be of independent interest. These empirical processes are weighted by the Radon-Nikodym derivatives mentioned earlier. Specifically, we

develop a Weighted Martingale Central Limit Theorem for functions of adaptively sampled data (Theorem C.5.1), as well as a novel Weighted Martingale Bernstein Inequality (Lemma C.7.2). Using these two results, we prove a functional asymptotic normality result for Radon-Nikodym derivative weighted empirical processes under bracketing number conditions. See Section 5.5.1 for more details.

5.2 Preliminaries

We consider a batch dataset collected by an adaptive sampling algorithm that pools across users. The dataset is comprised of data on n users over T decision times. For each decision time $t \in [1\colon T]$ and user $i \in [1\colon n]$, the observations consist of a multi-dimensional vector of random variables which we call the state, $S_t^{(i)} \in \mathbb{R}^{d_S}$; a scalar action (i.e., treatment), $A_t^{(i)} \in \mathcal{A}$ (here $\mathcal{A}$ is a finite set, so $|\mathcal{A}| < \infty$); and lastly the multi-dimensional outcome vector of random variables, $Y_t^{(i)} \in \mathbb{R}^{d_Y}$. Often adaptive sampling algorithms are designed to maximize a reward; in this case, the reward $R_t^{(i)} \in \mathbb{R}$ is some known function of the outcome vector $Y_t^{(i)}$. We define $Y_t^{(i)}$ because often we are interested in inference regarding quantities that are not the reward. For example, the reward in a physical activity digital health study could be the user's step count, but we may be interested in other outcomes like the user's heart rate.

We use potential outcomes[48] to represent counter-factual outcomes. We consider a longitudinal data setting in which the potential outcomes for $Y_t^{(i)}$ may depend on all actions taken on user i up to decision time t, $A_{1:t}^{(i)}$; we use the notation $A_{1:t}^{(i)} \triangleq \left\{A_{t'}^{(i)}\right\}_{t'=1}^{t}$ to denote collections of random variables. This means $Y_t^{(i)}$ has $|\mathcal{A}|^t$ different potential outcomes, $\left\{Y_t^{(i)}(a_{1:t}) : a_{1:t} \in \mathcal{A}^t\right\}$, where $\mathcal{A}^t$ denotes the t-fold Cartesian product of $\mathcal{A}$. Similarly, states have potential outcomes $\left\{S_t^{(i)}(a_{1:t-1}) : a_{1:t-1} \in \mathcal{A}^{t-1}\right\}$. The observed variables are $Y_t^{(i)} \triangleq Y_t^{(i)}\left(A_{1:t}^{(i)}\right)$ and $S_t^{(i)} \triangleq S_t^{(i)}\left(A_{1:t-1}^{(i)}\right)$.

We consider the setting in which the potential outcomes, $i \in [1\colon n]$, are i.i.d. according to an

unknown $\mathcal{P}$, i.e.,

$$D^{(i)} = \left\{ S_t^{(i)}(a_{1:t-1}), Y_t^{(i)}(a_{1:t}) : a_{1:t} \in \mathcal{A}^t \right\}_{t=1}^{T} \overset{i.i.d.}{\sim} \mathcal{P}, \text{ i.i.d over users } i \in [1\colon n]. \tag{5.2.1}$$

Note that the above allows for the trajectory of the observed user states and outcomes to be non-stationary and dependent over time. This setting encompasses both Markovian and non-Markovian user environments and is widely used in the longitudinal data analysis literature[35,83,84].

We consider adaptively sampled data for which at decision time $t = 1$, a pre-specified policy π_1, where $\mathbb{P}\left(A_1^{(i)} \middle| S_1^{(i)}\right) \triangleq \pi_1\left(A_1^{(i)}, S_1^{(i)}\right)$, is used to select the treatment actions independently for all users. Then, for each $t \geq 2$, an adaptive sampling algorithm may use all the observed data so far across all users to form a policy $\hat{\pi}_t^{(n)}$. Specifically the policy $\hat{\pi}_t^{(n)}$ can be formed using the history $\mathcal{H}_{t-1}^{(i)} \triangleq \left\{ S_{t'}^{(i)}, A_{t'}^{(i)}, Y_{t'}^{(i)} \right\}_{t'=1}^{t-1}$ for all users $i \in [1\colon n]$; for convenience we will use the notation $\mathcal{H}_{t-1}^{(1:n)} \triangleq \left\{ \mathcal{H}_{t-1}^{(i)} \right\}_{i=1}^{n}$ to represent the collective history for all users.

The policy $\hat{\pi}_t^{(n)}$ takes as input the user's current state, $S_t^{(i)}$, and outputs a sampling probability distribution over the action space $\mathcal{A}$. For $a \in \mathcal{A}$,

$$\hat{\pi}_t^{(n)}\left(a, S_t^{(i)}\right) = \mathbb{P}\left(A_t^{(i)} = a \middle| S_t^{(i)}, \mathcal{H}_{t-1}^{(1:n)}\right). \tag{5.2.2}$$

We consider data collected by adaptive sampling algorithms that, conditional on $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(1:n)}$, select actions $A_t^{(1)}, A_t^{(2)}, \ldots, A_t^{(n)}$ independently using policy $\hat{\pi}_t^{(n)}$. Note that the actions are not identically distributed conditional on $\mathcal{H}_{t-1}^{(1:n)}$ as the realized value of users' states at time t may differ.

Remark 5.2.1 (Dependent User Data Trajectories)**.** *Note that even though the users' potential outcomes $D^{(i)}$ are i.i.d. (as seen in display* (5.2.1)*), the observed user data trajectories, $\mathcal{H}_t^{(i)}$, are generally not independent over $i \in [1\colon n]$ due to algorithm's use of the common history, $\mathcal{H}_{t-1}^{(1:n)}$ in sampling the actions $A_t^{(1)}, A_t^{(2)}, \ldots, A_t^{(n)}$.*

For our statistical analyses, we consider asymptotics as the number of users, n, goes to infinity and keep the total number of decision times, T, fixed. This decision is motivated by our work in digital intervention experiments. These experiments are primarily concerned with using inference methods to draw scientific conclusions about a population of individuals over a fixed period of time, e.g., a 90-day physical activity mobile health intervention for individuals with stage-1 hypertension[67].

We now informally provide several key assumptions that we make on the online pooling adaptive sampling algorithm used to collect the data (see Section 5.3.2 for more details). We assume that policies $\hat{\pi}_t^{(n)}$ belong to a parametric class

$$\left\{\pi_t\big(\,\cdot\,;\beta_{t-1}\big) : \beta_{t-1} \in \mathbb{R}^{d_{t-1}}\right\}.$$

In particular, $\hat{\pi}_t^{(n)}(\,\cdot\,) \triangleq \pi_t\big(\,\cdot\,;\hat{\beta}t-1\big)$, where $\hat{\beta}t-1$ is a function of all users' data prior to time t, $\mathcal{H}_{t-1}^{(1:n)}$. For a given action $a \in \mathcal{A}$ and state $s \in \mathbb{R}^{d_S}$, $\pi_t\big(a,s;\hat{\beta}t-1\big)$ is a probability of selecting action a conditional on state s. We will assume conditions under which $\hat{\beta}t-1$ converges to a deterministic β_{t-1}^* as the number of users $n \to \infty$. Hence, we call $\pi_t(\,\cdot\,;\beta_{t-1}^*)$, which we abbreviate as π_t^*, the *target policy* at time t.

Remark 5.2.2 (Target Policies)**.** *Note that the assumption that there exists a target policy π_t^* for each decision time t is rather mild; this is because the asymptotic arguments derived here are as $n \to \infty$ with the total number of decision times T fixed. Further, β_{t-1}^* can be an arbitrary deterministic limit, e.g., β_{t-1}^* does not have to be a parameter in a correctly specified model of the reward, and π_t^* does not have to be optimal in any way. We also allow the target policy π_t^* to change with $t \in [2:\,T]$; this allows for non-stationarity in the users' outcomes that is not accounted for by the algorithm. In the special case that the environment is stationary and any models assumed by the algorithm are correctly specified, the target policy π_t^* may be the same for all $t \in [2:\,T]$. See Section 5.3.2 for more on the assumptions made*

on the policies.

5.3 Problem Statement

5.3.1 Inference Objective

We consider estimands that are defined with respect to the distribution in which the target policies $\pi^*_{2:T} \triangleq \{\pi^*_t\}_{t=2}^T$ are used to select actions. Specifically, we aim to conduct inference about a parameter θ^* that solves

$$0 = \mathbb{E}_{\pi^*_{2:T}}\left[\psi\big(\mathcal{H}_T^{(i)}; \theta\big)\right], \tag{5.3.1}$$

where ψ is a measurable function of $\mathcal{H}_t^{(i)}$ indexed by a finite-dimensional $\theta \in \mathbb{R}^{d_\theta}$.

The above expectation is indexed by the target policies $\pi^*_{2:T}$ to indicate that the expectation is over the distribution of $\mathcal{H}_T^{(i)}$ in which the actions are selected using the target policies and user potential outcomes are drawn from $\mathcal{P}$ as described in display (5.2.1); we will use $\mathcal{P}_{\pi^*}$ to refer to this distribution. Note that when the target policies $\pi^*_{2:T}$ are used to select actions, the data is no longer "adaptively sampled" so the user trajectories $\mathcal{H}_T^{(i)}$ are i.i.d.; thus, θ^* is not indexed by i.

To estimate θ^* we use Z-estimation; the estimator $\hat{\theta}$ satisfies

$$o_P\big(1/\sqrt{n}\big) = \frac{1}{n}\sum_{i=1}^{n} \psi\big(\mathcal{H}_T^{(i)}; \hat{\theta}\big). \tag{5.3.2}$$

This setup encompasses many types of standard estimators (e.g., least squares and maximum likelihood) and includes minimizers of differentiable loss functions. We are interested in constructing confidence regions for θ^*. We do this by characterizing the asymptotic distribution of $\hat{\theta}$ as the number of users $n \to \infty$ and using the asymptotic distribution to approximate the finite-sample distribution of $\hat{\theta}$.

To enhance expositional clarity we illustrate the ideas using the running example of a least squares

estimator in a binary action setting, $\mathcal{A} = \{0, 1\}$, with the following ψ:

$$\psi(\mathcal{H}_T^{(i)}; \theta) = \frac{1}{T} \sum_{t=1}^{T} \left(Y_t^{(i)} - \theta_0^\top S_t^{(i)} - A_t^{(i)} \theta_1^\top S_t^{(i)} \right) \begin{bmatrix} S_t^{(i)} \\ A_t^{(i)} S_t^{(i)} \end{bmatrix}. \tag{5.3.3}$$

Above, $Y_t^{(i)} \in \mathbb{R}$, $\theta = [\theta_0, \theta_1]$, and the first entry of $S_t^{(i)}$ is 1 (intercept term) for all t, i.

Remark 5.3.1 (Interpretation when the Model used by ψ is Misspecified)**.** *Often in Z-estimation, ψ corresponds to the derivative of a likelihood function for a parameter in a particular (possibly semi- or non-parametric) model for the data; in this case we can think of ψ as "correctly specified" if that model holds in our data. In the least squares example, ψ is correctly specified if $\mathbb{E}\left[Y_t^{(i)} \middle| S_t^{(i)}, A_t^{(i)}, \mathcal{H}_{t-1}^{(i)}\right] = \theta_0^{*,\top} S_t^{(i)} + A_t^{(i)} \theta_1^{*,\top} S_t^{(i)}$ a.s. for all t. As is standard for Z-estimators, if ψ is not correctly specified, then θ^* is the best* projected *solution; for example, in the least squares example from display* (5.3.3)*, θ^* corresponds to the best fitting linear model. The projection is with respect to the distribution $\mathcal{P}_{\pi^*}$ in which target policies $\pi^*_{2:T}$ are used to select actions. In this case θ^* is a function of the target policies $\pi^*_{2:T}$. However in the correctly specified model case, θ^* does not depend on the target policies. See Section 5.5.4 for how correct model specification affects the adaptive sandwich variance.*

Excursion Effects are a Key Use Case

Excursion effects, which are used for the primary analysis in micro-randomized trials [13,80,81], are a key use case for our inference method. In these longitudinal trials, treatment actions for each individual are repeatedly randomized using stochastic policies. The primary analysis for these trials concerns treatment effect excursions from the experiment's target policies. An example excursion effect is the following excursion from the target policy at time t:

$$\mathbb{E}_{\pi^*_{2:t-1}} \left[Y_t^{(i)}\left(A_{1:t-1}^{(i)}, a_t = 1\right) - Y_t^{(i)}\left(A_{1:t-1}^{(i)}, a_t = 0\right) \right]. \tag{5.3.4}$$

In the simplified setting in which the outcome $Y_t^{(i)}$ only depends on the most recent action $A_t^{(i)}$, the excursion effect simplifies to the standard treatment effect

$$\mathbb{E}\big[Y_t^{(i)}(a_t = 1) - Y_t^{(i)}(a_t = 0)\big].$$

Thus, the excursion effect from display (5.3.4) can be considered a generalization of the standard treatment effect to environments in which all actions taken so far, $A_{1:t}^{(i)}$, can affect the distribution of the outcome $Y_t^{(i)}$.

5.3.2 Policies Formed by the Adaptive Sampling Algorithm

As discussed in Section 5.2 (Preliminaries), at each decision time $t \in [2\colon T]$, the adaptive sampling algorithm uses all previously observed user data, $\mathcal{H}_{t-1}^{(1:n)}$, to form a policy $\hat{\pi}_t^{(n)}$ and uses this policy to select actions. In particular, we assume that there are policy function classes

$$\Big\{\pi_t(\,\cdot\,;\beta_{t-1}) : \beta_{t-1} \in \mathbb{R}^{d_{t-1}}\Big\} \tag{5.3.5}$$

for each $t \in [2\colon T]$ and that $\hat{\pi}_t^{(n)}(\,\cdot\,) \triangleq \pi_t(\,\cdot\,;\hat{\beta}t-1)$ where $\hat{\beta}t-1 \in \mathbb{R}^{d_{t-1}}$ is a statistic formed using $\mathcal{H}_{t-1}^{(1:n)}$.

Recall that we assume that for each decision time $t \in [2\colon T]$, the statistic $\hat{\beta}t-1$ formed by the algorithm converges in probability to a deterministic target policy parameter β_t^* as the number of users $n \to \infty$. The parameter β_t^* could parameterize a model of the expected reward used by the adaptive sampling algorithm. The target policy parameters $\{\beta_t^*\}_{t=1}^{T-1}$, which parameterize the target policies $\{\pi_t^*\}_{t=2}^{T} \triangleq \{\pi_t(\,\cdot\,;\beta_{t-1}^*)\}_{t=2}^{T}$. To allow for a large class of possible target policy parameters, we assume β_t^* is the solution to some estimating equation; we formally define these parameters below.

Recall that at the first decision time actions are selected using a pre-specified policy π_1. For the second decision time, the target policy $\pi_2^*(\,\cdot\,) \triangleq \pi_2(\,\cdot\,;\beta_1^*)$ where the target parameter β_1^* solves

$$0 = \mathbb{E}\left[\varphi_1\big(\mathcal{H}_1^{(i)};\beta_1^*\big)\right],$$

for some measurable function φ_1 of $\mathcal{H}_1^{(i)}$ indexed by a finite dimensional $\beta_1 \in \mathbb{R}^{d_1}$.

For the third decision time, the target policy $\pi_3^*(\,\cdot\,) \triangleq \pi_3(\,\cdot\,;\beta_2^*)$ where the target parameter β_2^* solves

$$0 = \mathbb{E}_{\pi_2^*}\left[\varphi_2\big(\mathcal{H}_2^{(i)};\beta_2^*\big)\right], \tag{5.3.6}$$

for some measurable function φ_2 of $\mathcal{H}_2^{(i)}$ indexed by a finite dimensional $\beta_2 \in \mathbb{R}^{d_2}$. In display (5.3.6) above, the expectation is indexed by target policies $\pi_2^*(\,\cdot\,)$. This means that the definition of the target policy for the third decision time, π_3^*, depends on the definition of target policy for the second decision time, π_2^*.

Continuing this pattern, for the t^{th} decision time, the target policy $\pi_t^*(\,\cdot\,) \triangleq \pi_t(\,\cdot\,;\beta_{t-1}^*)$ where the target parameter β_t^* solves

$$0 = \mathbb{E}_{\pi_{2:t-1}^*}\left[\varphi_{t-1}\big(\mathcal{H}_{t-1}^{(i)};\beta_{t-1}^*\big)\right], \tag{5.3.7}$$

for some measurable function φ_{t-1} of $\mathcal{H}_{t-1}^{(i)}$ indexed by a finite dimensional $\beta_{t-1} \in \mathbb{R}^{d_{t-1}}$. Again, the above expectation is indexed by the previous target policies $\pi_{2:t-1}^*$.

An example function φ_{t-1} is the following, which corresponds to a least squares solution:

$$\varphi_{t-1}\big(\mathcal{H}_{t-1}^{(i)};\beta_{t-1}\big) = \sum_{t'=1}^{t-1}\left(R_{t'}^{(i)} - \beta_{0,t-1}^\top S_{t'}^{(i)} - A_{t'}^{(i)}\beta_{1,t-1}^\top S_{t'}^{(i)}\right)\begin{bmatrix} S_{t'}^{(i)} \\ A_{t'}^{(i)} S_{t'}^{(i)}\end{bmatrix}. \tag{5.3.8}$$

In our simulations (Section 5.6) we consider a Boltzmann (or Softmax) exploration adaptive sam-

pling algorithm [9,20,92] that forms policies using estimators of the least squares solution β_t^*, defined with the estimating function from display (5.3.8).

Similarly, we assume that $\hat{\beta}t$, the estimators of the target policy parameters are Z-estimators, i.e., solutions to the empirical estimating functions. Formally, this means that $\hat{\beta}t - 1$ satisfies

$$o_P(1/\sqrt{n}) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{t-1}\left(\mathcal{H}_{t-1}^{(i)}; \hat{\beta}t - 1\right) \in \mathbb{R}^{d_{t-1}}. \tag{5.3.9}$$

Remark 5.3.2 (Misspecification of the Model used by the Adaptive Algorithm)**.** *In general, the adaptive sampling algorithms target parameter β_t^* defined in* (5.3.7) *can parameterize a model for parts of the multivariate distribution of $D^{(i)}$, the user's underlying potential outcomes from display* (5.2.1)*. We do not require that this model is correct. For example, even if the algorithm is developed assuming the environment corresponds to that of a stochastic contextual bandit, the validity of our statistical analysis will not be affected if this assumption is wrong.*

Key Assumptions on Policies

Rather than assume the adaptive sampling algorithm's model is correctly specified, we instead will make assumptions on the estimators $\left\{\hat{\beta}t\right\}_{t=1}^{T-1}$ of the policy parameters and the policy function classes $\left\{\pi_t\left(\cdot\,; \beta_{t-1}\right) : \beta_{t-1} \in \mathbb{R}^{d_{t-1}}\right\}$, from display (5.3.5). We now introduce the three foremost assumptions we place on the adaptive sampling policies, Conditions 5.3.1-5.3.3 below (we introduce the other assumptions we place on $\hat{\beta}t$ in Section 5.5.2).

Condition 5.3.1 is a consistency condition that ensures that the policy parameter estimator $\hat{\beta}t$ formed by the algorithm converges in probability to a target parameter value β_t^* as the number of users $n \to \infty$.

Condition 5.3.1 (Consistency of Policy Estimators). *For each* $t \in [1 : T-1]$,

$$\hat{\beta}t \xrightarrow{p} \beta_t^*.$$

Remark 5.3.3 (Sufficient Assumptions for Condition 5.3.1). *In Theorem C.2.1 of Appendix C.2 we state simple sufficient conditions for Condition 5.3.1 to hold.*

The next two key assumptions we place on the adaptive sampling algorithm, Conditions 5.3.2 and 5.3.3, both concern the policy function classes $\{\pi_t(\cdot\,; \beta_{t-1}) : \beta_{t-1} \in \mathbb{R}^{d_{t-1}}\}$, from display (5.3.5).

Condition 5.3.2 (Minimum Exploration). *Let* $0 < \pi_{\min} < 1$ *be a constant. For all* $t \in [2 : T]$,

$$\inf_{\beta_{t-1} \in \mathbb{R}^{d_{t-1}}} \pi_t(a, s; \beta_{t-1}) \geq \pi_{\min}$$

for all $a \in \mathcal{A}$ *and* $s \in \mathcal{S}$. *Also for* $t = 1$, $\pi_1(a, s) \geq \pi_{\min}$ *for all* $a \in \mathcal{A}$ *and* $s \in \mathcal{S}$.

Condition 5.3.2 ensures that the policy class produces action selection probabilities that are strictly bounded above zero for all actions. Note this ensures that the policy is stochastic, as is necessary for micro-randomized trials (discussed earlier in Section 5.3.1). Note this condition excludes deterministic policies, which means target policies that maximize the expected reward in classical contextual bandit and Markov decision process environments are excluded. However, in general, the fewer structural assumptions that are placed on the environment, the more need there is for reward-maximizing algorithms to continually explore. For example, in non-stationary and adversarial sequential decision-making problem settings it is common both theoretically and in practice to prevent the RL algorithm's action selection probabilities to go to zero for any action [16,21,22,63] in order to ensure the algorithm can detect changes in the reward distribution. Action selection proba-

bilities are also commonly constrained away from 0 and 1 to facilitate causal inference and off-policy evaluation after the experiment is over[39,68,94,96,104].

In Condition 5.3.3 below, for each $t \in [1\colon T-1]$, we use $B_t \subset \mathbb{R}^{d_t}$ to denote some compact subset whose interior contains β_t^*.

Condition 5.3.3 (Lipschitz Policy Functions). *For all $t \in [2\colon T]$, there is a non-negative, real-valued function $\dot{\pi}_t(A_t^{(i)}, S_t^{(i)})$ such that (i) $\mathbb{E}_{\pi_{2:t}^*}\left[\left|\dot{\pi}_t(A_t^{(i)}, S_t^{(i)})\right|^{2+\alpha}\right] < \infty$ for some $\alpha > 0$, and (ii) for any $\beta_{t-1}, \beta'_{t-1} \in B_{t-1}$,*

$$\left|\pi_t(A_t^{(i)}, S_t^{(i)}; \beta_{t-1}) - \pi_t(A_t^{(i)}, S_t^{(i)}; \beta_{t-1})\right| \leq \dot{\pi}_t(A_t^{(i)}, S_t^{(i)})\left\|\beta_{t-1} - \beta'_{t-1}\right\|_2 \text{ a.s.}$$

Condition 5.3.3 is a smoothness condition on the policy function classes, which excludes policies that are a discontinuous function of parameters β_{t-1}. It is well known in the inference after adaptive sampling literature that standard estimators, like the sample mean, can be asymptotically non-normal on data collected by adaptive algorithms that do not satisfy such smoothness conditions[26,39,109]. Although this smoothness condition may appear rather mild, note that the reward-maximizing policy in a stochastic bandit problem is a discontinuous function of the margin because of the argmax operation; for example, in a two-armed bandit setting with $\beta^* \triangleq \mathbb{E}[R_t(1)] - \mathbb{E}[R_t(0)]$, the optimal policy is $\mathbb{P}(A_t = 1) = \mathbb{I}_{\beta^* > 0}$. Despite this, as mentioned after Condition 5.3.2, there are standard reinforcement learning algorithms developed for more complex environments (e.g., non-stationary) which satisfy this smoothness condition.

Remark 5.3.4 (Example Algorithms that Satisfy Conditions 5.3.2 and 5.3.3). *In Appendix C.1.2 we show that a Boltzmann (or Softmax) exploration algorithm[9,20,92] and a stochastic mirror descent algorithm (based on those from[63] pg 361 and[16]) both satisfy Conditions 5.3.2 and 5.3.3 above.*

5.4 Related Work

Recently, many inference methods have been developed for adaptively sampled data focused on multi-armed and contextual bandit environments. These include inference methods via asymptotic approximations [11,23,26,39,108,109,110] as well as approaches that use high probability bounds [2,15,43,54]. These works for the most part consider asymptotics as $T \to \infty$. These methods are more restrictive than ours in that they assume an underlying contextual bandit environment that does not allow a user's potential outcomes to be dependent over time. Moreover, most of these approaches consider inference for particular estimands, e.g., the value of a policy or a specific treatment effect, rather than an all-purpose Z-estimand. However, these methods are more general than ours in that they put fewer restrictions on the adaptive sampling policies used to collect the batch data, e.g., many allow the action selection probabilities to go to zero at some rate for some actions and do not require their policy function classes to be smooth in its parameters. Additionally, most of these prior methods require that the reward model used by the adaptive sampling algorithm is correctly specified.

The adaptive clinical trial literature provides methods for inference after using policies that satisfy conditions akin to Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) above, e.g., see Theorems 3.1 and 9.1 of [44]. There are two ways in which our results differ from these classical results. The first is that we consider a setting in which the adaptive sampling algorithm repeatedly selects treatment actions for each of multiple individuals sequentially over time. Since the adaptive sampling algorithm selects actions probabilistically, each individual is sequentially randomized. In contrast, the adaptive clinical trial literature classically considers settings in which at each decision time a new individual is drawn independently from the population and the adaptive sampling algorithm makes one treatment action decision per individual. The second major difference is that these classical results assume that both the model used for inference and the model used by the adaptive algorithm are correctly specified. For our results, we do not assume either of these models

is correctly specified; see Remark 5.3.1 for more on model misspecification. Additionally, in Section 5.5.4 we discuss how our asymptotic results simplify when the estimating function, ψ uses a correctly specified model.

Another area of related work is inference methods for longitudinal data. This literature assumes the same underlying potential outcomes model, display (5.2.1), that allows for non-stationarity and dependent outcomes over time within each user [35,84,106]. However, this literature considers batch datasets in which user data trajectories are independent across users, which excludes datasets collected by adaptive sampling algorithms that learn across users. This literature also includes methods for inferring excursion effects [13,81], which were discussed in Section 5.3.1.

Here we generalize techniques from the classical literature on empirical processes for i.i.d. data [99] to adaptively sampled data. In particular, we develop functional asymptotic normality and maximal inequality results for "Radon-Nikodym derivative weighted empirical processes"; see Section 5.5.1 for more details. Note that [12] develops a maximal inequality for adaptively sampled data assuming a classical contextual bandit environment. Besides the differences in the underlying environment assumptions, our maximal inequality results also differ from theirs because they consider asymptotics as $T \to \infty$, while here T is fixed and we consider asymptotics as $n \to \infty$.

5.5 Main Results

Note that if the batch data were collected using the fixed target policies $\pi^*_{2:T}$, rather than the data-dependent, adaptive policies $\hat{\pi}^{(n)}_{2:T}$, then the data trajectories would be independent across users, i.e., $\mathcal{H}_T^{(i)}$ would be i.i.d. across $i \in [1\colon n]$. In that i.i.d. setting, we could use standard asymptotic normality results for Z-estimators [99] Theorem 5.21 to get that $\hat{\theta}$ is asymptotically normal with the

standard sandwich variance, i.e.,

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{D} \mathcal{N}\left(0,\ [\dot{\Psi}^*]^{-1}\Sigma[\dot{\Psi}^*]^{-1,\top}\right), \tag{5.5.1}$$

with "bread" $\dot{\Psi}^* \triangleq \frac{\partial}{\partial\theta}\mathbb{E}_{\pi^*_{2:T}}\left[\psi\left(\mathcal{H}_T^{(i)};\theta\right)\right]\big|_{\theta=\theta^*}$ and "meat" $\Sigma \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\psi(\mathcal{H}_T^{(i)};\theta^*)^{\otimes 2}\right]$; we use the notation $x^{\otimes 2} \triangleq xx^\top$.

However, in the adaptively sampled data setting in which the random, data-dependent policies $\hat{\pi}^{(n)}_{2:T}$ produced by the adaptive sampling algorithm are used to select actions, we show that the limiting variance is different, specifically,

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{D} \mathcal{N}\left(0,\ [\dot{\Psi}^*]^{-1}\Sigma^{\text{adapt}}[\dot{\Psi}^*]^{-1,\top}\right), \tag{5.5.2}$$

where

$$\Sigma^{\text{adapt}} \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\left\{\psi\left(\mathcal{H}_T^{(i)};\theta^*\right) + \dot{\Psi}^*\sum_{t=1}^{T-1} M_t\,\varphi_t\left(\mathcal{H}_t^{(i)};\beta_t^*\right)\right\}^{\otimes 2}\right]. \tag{5.5.3}$$

Above $M_t \in \mathbb{R}^{d_\theta \times d_t}$. See display (5.5.32) for the definition of the M_t matrices.

We call the limiting variance in display (5.5.2), the *adaptive* sandwich variance. Comparing Σ and Σ^{adapt} from display (5.5.3), we can interpret the term $\dot{\Psi}^*\sum_{t=1}^{T-1} M_t\,\varphi_t\left(\mathcal{H}_t^{(i)};\beta_t^*\right)$ as the "cost" or "inflation" in variance due to using the estimated $\hat{\pi}^{(n)}_{2:T}$ to select actions rather than $\pi^*_{2:T}$. In special cases, under a property we call "policy invariance", $M_t = 0$ so the limiting sandwich and adaptive sandwich variances are equal; see Section 5.5.4 for more details. We provide estimators of the adaptive sandwich variance (see Appendix C.1.1), which we use in Section 5.6 for our simulation results.

We now outline the remainder of this asymptotic results section. In order to provide a high-level understanding of how our proof techniques and results differ from the i.i.d. data case, in Section 5.5.1 we discuss the ideas behind our asymptotic normality proof; specifically, we introduce the Radon-Nikodym derivative weighting we use in Section 5.5.1 and provide an overview of the func-

tional asymptotic normality results for adaptively sampled data that we develop in Section 5.5.1. Then in Section 5.5.2, we state our results formally; specifically in Section 5.5.2 we introduce additional assumptions on the policy parameters and in Section 5.5.2 we have our formal theorem statements. In Section 5.5.3, we provide a more detailed proof sketch of our main asymptotic normality result, display (5.5.2). Finally, in Section 5.5.4, we discuss cases in which the limiting adaptive sandwich variance equals the standard sandwich variance.

5.5.1 Ideas Behind the Proof of Asymptotic Normality

In order to provide a high-level understanding of how proving results for adaptively sampled data differs from the i.i.d. data case, we now discuss the key ideas that we use in our proof of asymptotic normality. The foremost technical challenge in our proof of asymptotic normality of $\sqrt{n}\big(\hat{\theta} - \theta^*\big)$ is to account for how the data is collected using estimated policies $\hat{\pi}_t^{(n)}(\,\cdot\,) = \pi_t\big(\,\cdot\,; \hat{\beta}t - 1\big)$. Specifically, the challenge is accounting for how the error of estimator $\hat{\theta}$ is impacted by the error of the estimated policies $\hat{\pi}_t^{(n)}(\,\cdot\,) = \pi_t\big(\,\cdot\,; \hat{\beta}t - 1\big)$ used to collect the data.

A key insight of this work is that *a Z-estimator $\hat{\theta}$ formed on adaptively sampled data can be framed as a Z-estimator in which the estimated policy parameters used to collect the data, $\hat{\beta}1 : T - 1$, are plug-in estimates of nuisance parameters.* Classically on i.i.d. data, one constructs Z-estimators with a plug-in estimate of nuisance parameters which are fitted on the same data as the one used to form the Z-estimator itself[98]. In deriving the asymptotic results for the Z-estimator in these classical settings, one must account for the dependence between the Z-estimator of interest and the plug-in estimator because they are constructed using a shared dataset.

However, how to frame the inference after adaptive sampling problem as a problem of inference via a Z-estimator with a plug-in nuisance parameter estimator does not follow straightforwardly from the classical literature. Recall from the definition of $\hat{\theta}$ from display (5.3.2) that $\hat{\theta}$ is formed by the data analyzer without constructing any plug-in estimators for nuisance parameters. We are

interested in using the methods from the literature on plug-in estimates to account for the error of the policy parameters $\hat{\beta}1:T-1$, which impacted how the data was *collected*. On i.i.d. data though, plug-in estimators do not affect *data collection* and are only used for the *data analysis*.

The critical step that will allow us to treat the policy parameters as plug-in nuisance parameters is to write the estimating function for the Z-estimator $\hat{\theta}$ and the policy parameters $\hat{\beta}1:T-1$ jointly. The key tool we will use to do this is Radon-Nikodym derivative weighting.

Radon-Nikodym Derivative Weights

Note the following ratios:

$$\frac{\pi_t\big(\,\cdot\,, S_t^{(i)}; \beta_{t-1}\big)}{\hat{\pi}_t^{(n)}\big(\,\cdot\,, S_t^{(i)}\big)} : \mathcal{A} \mapsto [0, \infty). \tag{5.5.4}$$

Conditional on $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$, the functions $\hat{\pi}_t^{(n)}\big(\,\cdot\,, S_t^{(i)}\big)$ and $\pi_t\big(\,\cdot\,, S_t^{(i)}; \beta_{t-1}\big)$ each define a probability distribution over the action space $\mathcal{A}$. The ratio of these two probability distributions, as seen in display (5.5.4) above, is a Radon-Nikodym derivative; see Lemma C.1.3 for a formal statement of this result. Note that in the proof of Lemma C.1.3, we use the minimum exploration Condition 5.3.2 (Minimum Exploration), to ensure that these Radon-Nikodym derivatives exist.

For notational convenience, we define the following weighting functions for any $\beta_{t-1}, \beta'_{t-1} \in \mathbb{R}^{d_{t-1}}$,

$$W_t^{(i)}(\beta_{t-1}, \beta'_{t-1}) \triangleq \frac{\pi_t\big(A_t^{(i)}, S_t^{(i)}; \beta_{t-1}\big)}{\pi_t\big(A_t^{(i)}, S_t^{(i)}; \beta'_{t-1}\big)}. \tag{5.5.5}$$

Additionally, for any $\beta_{1:t-1}, \beta'_{1:t-1} \in \mathbb{R}^{d_{1:t-1}}$ (where $d_{1:t-1} \triangleq \sum_{t'=1}^{t-1} d_{t'}$) we define,

$$W_{2:t}^{(i)}(\beta_{1:t-1}, \beta'_{1:t-1}) \triangleq \prod_{t'=2}^{t} W_{t'}^{(i)}(\beta_{t'-1}, \beta'_{t'-1}).$$

The Radon-Nikodym weights above allow us to define estimating functions such as

$$\Psi\left(\beta_{1:T-1}, \theta\right) \triangleq \mathbb{E}_{\pi(\beta_{1:T-1})}\left[\psi\left(\mathcal{H}_T^{(i)}; \theta\right)\right] = \mathbb{E}\left[W_{2:T}^{(i)}(\beta_{1:T-1}, \hat{\beta}1 : T-1)\psi\left(\mathcal{H}_T^{(i)}; \theta\right)\right]. \tag{5.5.6}$$

Above we use $\mathbb{E}$ to denote expectations with respect to the distribution used to collect the data. Thus, in the expectations from displays (5.5.6) and (5.5.8) above, the Radon-Nikodym weights have the effect of changing the distribution the expectation is taken over. Specifically, it changes the policy with which the actions are selected. Above we use the notation $\mathbb{E}_{\pi(\beta_{1:T-1})}$ to denote the expectation with respect to the distribution in which (i) the policies $\left\{\pi_t(\,\cdot\,; \beta_{t-1})\right\}_{t=2}^{T}$ are used to select actions and (ii) user potential outcomes are drawn from $\mathcal{P}$, as described in display (5.2.1).

We also define an empirical version of the limiting estimating function $\Psi\left(\beta_{1:T-1}, \theta\right)$ above:

$$\hat{\Psi}^{(n)}\left(\beta_{1:T-1}, \theta\right) \triangleq \frac{1}{n}\sum_{i=1}^{n} W_{2:T}^{(i)}(\beta_{1:T-1}, \hat{\beta}1 : T-1)\psi\left(\mathcal{H}_T^{(i)}; \theta\right). \tag{5.5.7}$$

Additionally, in our proofs we will use the following limiting and empirical estimating functions for the policy parameters:

$$\Phi_{1:T-1}(\beta_{1:T-1}) \triangleq \mathbb{E}\begin{bmatrix} \varphi_1(\mathcal{H}_1^{(i)}; \beta_1) \\ W_2^{(i)}(\beta_1, \hat{\beta}1)\varphi_2(\mathcal{H}_2^{(i)}; \beta_2) \\ W_{2:3}^{(i)}(\beta_{1:2}, \hat{\beta}1 : 2)\varphi_3(\mathcal{H}_3^{(i)}; \beta_3) \\ \vdots \\ W_{2:T-1}^{(i)}(\beta_{1:T-2}, \hat{\beta}1 : T-2)\varphi_{T-1}(\mathcal{H}_{T-1}^{(i)}; \beta_{T-1}) \end{bmatrix} \tag{5.5.8}$$

$$\hat{\Phi}^{(n)}_{1:T-1}(\beta_{1:T-1}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1) \\ W_2^{(i)}(\beta_1,\hat{\beta}1)\varphi_2(\mathcal{H}_2^{(i)};\beta_2) \\ W_{2:3}^{(i)}(\beta_{1:2},\hat{\beta}1:2)\varphi_3(\mathcal{H}_3^{(i)};\beta_3) \\ \vdots \\ W_{2:T-1}^{(i)}(\beta_{1:T-2},\hat{\beta}1:T-2)\varphi_{T-1}(\mathcal{H}_{T-1}^{(i)};\beta_{T-1}) \end{bmatrix} \tag{5.5.9}$$

We use these joint estimating functions for both the policy parameters $\beta_{1:T-1}$ and the parameter of interest θ to derive the joint limiting distribution of $\hat{\beta}1:T-1$ and $\hat{\theta}$. Specifically, we prove that $\hat{\beta}1:T-1$ and $\hat{\theta}$ are jointly asymptotically normal, as seen in display (5.5.10) below; see Section 5.5.3 for a proof sketch. Note that since our adaptively sampled data is non-i.i.d. the proof of this result relies heavily on novel functional asymptotic normality results for Radon-Nikodym weighted functions on adaptively sampled data, which we discuss in detail in Section 5.5.1.

$$\sqrt{n}\begin{pmatrix} \hat{\beta}1:T-1-\beta^*_{1:T-1} \\ \hat{\theta}-\theta^* \end{pmatrix} \xrightarrow{D} \mathcal{N}\left(0, \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} \Sigma_{1:T} \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1,\top}\right). \tag{5.5.10}$$

Above,

$$\Sigma_{1:T} \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\begin{pmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1^*) \\ \varphi_2(\mathcal{H}_2^{(i)};\beta_2^*) \\ \vdots \\ \varphi_{T-1}(\mathcal{H}_{T-1}^{(i)};\beta_{T-1}^*) \\ \psi(\mathcal{H}_T^{(i)};\theta^*) \end{pmatrix}^{\otimes 2}\right]. \tag{5.5.11}$$

and

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix} \triangleq \begin{bmatrix} \frac{\partial}{\partial \beta_{1:T-1}} \Phi_{1:T-1}(\beta_{1:T-1}) & \frac{\partial}{\partial \theta} \Phi_{1:T-1}(\beta_{1:T-1}) \\ \frac{\partial}{\partial \beta_{1:T-1}} \Psi(\beta_{1:T-1}, \theta) & \frac{\partial}{\partial \theta} \Psi(\beta_{1:T-1}, \theta) \end{bmatrix} \Big|_{(\beta_{1:T-1}, \theta) = (\beta^*_{1:T-1}, \theta^*)}. \tag{5.5.12}$$

Note above that $\Phi_{1:T-1}(\beta^*_{1:T-1})$ is not a function of θ, thus $\frac{\partial}{\partial \theta} \Phi_{1:T-1}(\beta^*_{1:T-1})\big|_{\theta=\theta^*} = 0$.

Display (5.5.10) above is sufficient for showing that $\sqrt{n}\big(\hat{\theta} - \theta^*\big)$ is asymptotically normal with the adaptive sandwich variance from display (5.5.1) holds. Specifically, by Lemma C.3.1 (Equivalent Formulations for the Adaptive Sandwich Variance), the lower $d_\theta \times d_\theta$ block of the limiting variance matrix from display (5.5.10) above is equivalent to the adaptive sandwich variance $[\dot{\Psi}^*]^{-1} \Sigma^{\text{adapt}} [\dot{\Psi}^*]^{-1,\top}$ from display (5.5.2).

Overview of Functional Asymptotic Normality Result for Adaptively Sampled Data

The proof of the asymptotic normality result from display (5.5.10) relies heavily on a novel functional asymptotic normality results we develop for adaptively sampled data. Functional asymptotic normality results are classical results from the empirical process literature that are used in many Z-estimator asymptotic normality proofs. These classical results concern stochastic processes of the following form:

$$\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(f(\mathcal{H}_T^{(i)}) - \mathbb{E}\big[f(\mathcal{H}_T^{(i)})\big] \Big) \;\; \text{s.t. } f \in \mathcal{F} \right\}, \tag{5.5.13}$$

for a class of functions $\mathcal{F}$. If user data trajectories $\mathcal{H}_T^{(i)}$ were i.i.d. over $i \in [1\colon n]$ and the complexity of a class of real-valued functions $\mathcal{F}$ was properly controlled, then the stochastic process from display (5.5.13) would converge in distribution to a Gaussian process; see Theorem 19.5 of[99].

For our inference after adaptive sampling problem, we are interested in showing a functional

asymptotic normality result for stochastic processes like the following:

$$\left\{\sqrt{n}\left[\hat{\Psi}^{(n)}(\beta_{1:T-1},\theta)-\Psi(\beta_{1:T-1},\theta)\right] \text{ s.t. } \beta_{1:T-1}\in B_{1:T-1},\theta\in\Theta\right\}$$
$$=\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(W_{2:T}^{(i)}(\beta_{1:T-1},\hat{\beta}1:T-1)\psi(\mathcal{H}_T^{(i)};\theta)-\mathbb{E}\left[W_{2:T}^{(i)}(\beta_{1:T-1},\hat{\beta}1:T-1)\psi(\mathcal{H}_T^{(i)};\theta)\right]\right)\right.$$
$$\left.\text{s.t. } \beta_{1:T-1}\in B_{1:T-1},\ \theta\in\Theta\right\}. \quad (5.5.14)$$

Above $B_{1:T-1}\subset\mathbb{R}^{d_{1:T-1}}$ and $\Theta\subset\mathbb{R}^{d_\theta}$ are compact balls whose interiors contain $\beta_{1:T-1}^*$ and θ^* respectively. Also, recall that above we use the expectation $\mathbb{E}$ (not indexed by any policies) to refer to the distribution which was used to generate the data. Note that this means that

$\mathbb{E}\left[W_{2:T}^{(i)}(\beta_{1:T-1},\hat{\beta}1:T-1)\psi(\mathcal{H}_T^{(i)};\theta)\right]=\mathbb{E}_{\pi(\beta_{1:T-1})}\left[\psi(\mathcal{H}_T^{(i)};\theta)\right]$.

Note that to show that the stochastic process from display (5.5.14) is functionally asymptotically normal, it is sufficient to show a functional asymptotic normality result for empirical processes of the form

$$\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\{\hat{\pi}_{2:T}^{(i)}\}^{-1}f(\mathcal{H}_T^{(i)})-\mathbb{E}\left[\{\hat{\pi}_{2:T}^{(i)}\}^{-1}f(\mathcal{H}_T^{(i)})\right]\right) \text{ s.t. } f\in\mathcal{F}\right\}, \quad (5.5.15)$$

where $\hat{\pi}_{2:T}^{(i)}\triangleq\prod_{t=2}^{T}\hat{\pi}_t\left(A_t^{(i)},S_t^{(i)}\right)$. Specifically, the stochastic process from display (5.5.14) is equivalent to the stochastic process in display (5.5.15) for the following choice of $\mathcal{F}$:

$$\mathcal{F}=\left\{\left[\prod_{t=2}^{t}\pi_t(\,\cdot\,;\beta_{t-1})\right]\psi(\,\cdot\,;\theta) \text{ s.t. } \beta_{1:T-1}\in B_{1:T-1},\theta\in\Theta\right\}.$$

By Theorem 18.14 of[99], the two conditions needed to ensure a functional normality result for display (5.5.15) are (i) a joint asymptotic normality result for the stochastic process evaluated at any finite number of functions $f_1,f_2,\ldots,f_k\in\mathcal{F}$, and (ii) a maximal inequality result over the function class $\mathcal{F}$. To show part (i) holds for adaptively sampled data, we prove a Weighted Martingale

Triangular Array Central Limit Theorem (Theorem C.5.1). Specifically this Theorem can be used to show asymptotic normality results like the following:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\{\hat{\pi}_{2:T}^{(i)}\}^{-1}f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\{\hat{\pi}_{2:T}^{(i)}\}^{-1}f(\mathcal{H}_T^{(i)})\right]\right) \\ \xrightarrow{D} \mathcal{N}\left(0,\ \mathbb{E}_{\pi^*_{2:T}}\left[\{\pi_{2:T}^{*,(i)}\}^{-2}f(\mathcal{H}_T^{(i)})^2\right] - \mathbb{E}_{\pi^*_{2:T}}\left[\{\pi_{2:T}^{*,(i)}\}^{-1}f(\mathcal{H}_T^{(i)})\right]^2\right).$$

Above $\pi_{2:T}^{*,(i)} \triangleq \prod_{t=2}^{T}\pi_t^*\left(A_t^{(i)}, S_t^{(i)}\right)$. The proof of our asymptotic normality result heavily relies on Lipschitz policy function Condition 5.3.3 and builds on the martingale Central Limit Theorem from Theorem 2.2 of[29].

For part (ii), we prove a maximal inequality for adaptively sampled data as a function of the bracketing integral of $\mathcal{F}$, Lemma C.8.1. We prove our maximal inequality, Lemma C.8.1, using a novel Weighted Martingale Bernstein Inequality, Lemma C.7.2. This inequality modifies the classical Bernstein inequality for i.i.d. data[99] Lemma 19.32. Specifically, our Bernstein inequality ensures that on our adaptively sampled data type, for any real-valued function f of $\mathcal{H}_T^{(i)}$ with $\|f\|_\infty < \infty$,

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\hat{\pi}_{2:T}^{(i)}\}^{-1}f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\{\hat{\pi}_{2:T}^{(i)}\}^{-1}f(\mathcal{H}_T^{(i)})\right]\right| > x\right) \\ \leq 2\exp\left(-\frac{\pi_{\min}^{T-1}}{4}\frac{x^2}{\mathbb{E}_{\pi^*_{2:T}}\left[\{\pi_{2:T}^{*,(i)}\}^{-1}f(\mathcal{H}_T^{(i)})^2\right] + x\|f\|_\infty/\sqrt{n}}\right), \quad (5.5.16)$$

for any $x > 0$ and $n \geq 1$.

We now discuss the key techniques used in the proof of Lemma C.7.2. Our proof of Lemma C.7.2, similar to the classical Bernstein inequality proof, starts by using a Chernoff bound to get an upper tail bound. Specifically, for any $\lambda > 0$,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\{\hat{\pi}_{2:T}^{(i)}\}^{-1}f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\{\hat{\pi}_{2:T}^{(i)}\}^{-1}f(\mathcal{H}_T^{(i)})\right]\right) > x\right)$$

$$\leq e^{-\lambda x}\mathbb{E}\left[\exp\left\{\frac{\lambda}{\sqrt{n}}\sum_{i=1}^{n}\left(\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)})\right]\right)\right\}\right].$$

Changing the summation in exponent into a product,

$$= e^{-\lambda x}\mathbb{E}\left[\prod_{i=1}^{n}\exp\left\{\frac{\lambda}{\sqrt{n}}\left(\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)})\right]\right)\right\}\right].$$

In the original proof for i.i.d. data, the next step is to move the product over $i = 1, 2, \ldots n$ above outside of the expectation. If we omit the terms $\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}$ above and if user trajectories $\mathcal{H}_T^{(i)}$ were i.i.d. over $i \in [1\colon n]$, moving the product outside the expectation would be trivial. However, this is not the case for our adaptively sampled data setting.

The key insight we use in our proof is Lemma C.7.1, a result that allows us to move products out of expectations using the $\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}$ weighting. Specifically, this Lemma proves that for any constant c,

$$\mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)}) + c\right)\right] = \prod_{i=1}^{n}\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)}) + c\right].$$

The proof leverages the conditional independence of the action selection at each time step and the fact that the underlying potential outcomes are i.i.d. See Appendix C.7 for more details on all our maximal inequality results.

5.5.2 Formal Statement of Results

We now formally state the additional conditions we use to show consistency and asymptotic normality of $\hat{\theta}$. Below we first provide assumptions on the estimated policy parameters $\hat{\beta}1 : T-1$ (Section 5.5.2) and then provide the theorem for, and assumptions on, the estimator $\hat{\theta}$ based on the resulting adaptively sampled data (Section 5.5.2). The rationale for this is first, separating out the conditions on the inferential estimator $\hat{\theta}$ will also make explicit the conditions placed on the $\hat{\theta}$ Z-estimator due to the adaptive sampling. A second consideration is that designers of adaptive sam-

pling algorithms will know what assumptions on the algorithm are sufficient so that the resulting data can be used in a wide variety of after-study data analyses. A third consideration is that a data analyst who is provided an adaptively sampled dataset (with a known algorithm) can devise tests on the data to challenge the assumptions made on the policy parameters $\hat{\beta}1 : T-1$.

Assumptions on the Policy Parameters

We now discuss formally the remaining assumptions we place on the policy parameter estimators $\hat{\beta}t$ and their estimating functions φ_t. (Recall the assumption that estimators $\hat{\beta}t$ are consistent for β_t^* and the assumptions placed on the policy function classes $\left\{\pi_t(\,\cdot\,;\beta_{t-1}) : \beta_{t-1} \in \mathbb{R}^{d_{t-1}}\right\}$ were introduced earlier in Section 5.3.2.) The first of these assumptions, Condition 5.5.1 below, will use the notation $\mathbb{E}_{\pi(\beta_{1:T-1})}$ to denote the expectation with respect to the distribution in which (i) the policies $\left\{\pi_t(\,\cdot\,;\beta_{t-1})\right\}_{t=2}^{T}$ are used to select actions and (ii) user potential outcomes are drawn from $\mathcal{P}$, as described in display (5.2.1).

Condition 5.5.1 (Differentiability of Policy Parameter Estimating Functions)**.** *The following mapping is differentiable at* $\beta_{1:T-1} = \beta^*_{1:T-1}$

$$\beta_{1:T-1} \mapsto \Phi_{1:T-1}(\beta_{1:T-1}) \triangleq \mathbb{E}_{\pi(\beta_{1:T-1})}\begin{bmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1) \\ \varphi_2(\mathcal{H}_2^{(i)};\beta_2) \\ \vdots \\ \varphi_{T-1}(\mathcal{H}_{T-1}^{(i)};\beta_{T-1}) \end{bmatrix}. \tag{5.5.17}$$

Above the function $\Phi_{1:T-1}(\beta_{1:T-1})$ *was first defined in display* (5.5.17)*. We also assume that for each* $t \in [1 : T-1]$*, the derivative matrix* $\dot{\Phi}_t^* \triangleq \frac{\partial}{\partial \beta_t}\mathbb{E}_{\pi^*_{2:t}}\left[\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right]\big|_{\beta_t=\beta_t^*}$ *is invertible.*

Note in the expectation above in display (5.5.17) *that* $\mathcal{H}_t^{(i)}$ *only depends on the policies used to select actions up to decision time t, i.e., policies* $\left\{\pi_{t'}(\,\cdot\,;\beta_{t'-1})\right\}_{t'=2}^{t}$*. Thus,* $\mathbb{E}_{\pi(\beta_{1:T-1})}\left[\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right] =$

$\mathbb{E}_{\pi(\beta_{1:t-1})}\left[\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right]$.

Condition 5.5.1 ensures that the estimating functions for the policy parameters $\beta_{1:T-1}$ are differentiable at $\beta_{1:T-1} = \beta^*_{1:T-1}$. This kind of differentiability condition is common for Z-estimators [99] Theorem 5.21. What is notable about Condition 5.5.1 is that in display (5.5.17), the policy parameters $\beta_{1:T-1}$ parameterize not only the estimating functions $\varphi_t(\mathcal{H}_t^{(i)};\beta_t)$, they also parameterize the distribution with which the expectation is taken, $\mathbb{E}_{\pi(\beta_{1:T-1})}$.

The final assumption we place on the policies, Condition 5.5.2, is a Lipschitz condition on the policy parameter estimating functions φ_t. This condition has the effect of restricting the complexity of the function class $\left\{\varphi_t(\,\cdot\,;\beta_t) \,:\, \beta_t \in B_t\right\}$ and has been used in other standard proofs for the asymptotic normality of Z-estimators, e.g., see Theorem 5.21 of [99]. Below, for each $t \in [1\colon T-1]$, we use $B_t \subset \mathbb{R}^{d_t}$ to denote some compact subset whose interior contains β^*_t.

Condition 5.5.2 (Lipschitz Policy Estimating Function). *Let $\alpha > 0$ be a constant. For each $t \in [2\colon T]$, there is a non-negative valued function $\dot{\varphi}_t(\mathcal{H}_t^{(i)})$ such that*

(i) For any $\beta_t, \beta'_t \in B_t$,

$$\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t) - \varphi_t(\mathcal{H}_t^{(i)};\beta'_t)\right\|_2 \le \dot{\varphi}_t(\mathcal{H}_t^{(i)})\left\|\beta_t - \beta'_t\right\|_2 \text{ a.s.} \tag{5.5.18}$$

(ii) $\mathbb{E}_{\pi^*_{2:t}}\left[\left|\dot{\varphi}_t(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right] < \infty$.

(iii) $\mathbb{E}_{\pi^*_{2:t}}\left[\left|\dot{\varphi}_t(\mathcal{H}_t^{(i)})\dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)})\right|^{2+\alpha}\right] < \infty$ *for all $t' \in [2\colon t]$ (the function $\dot{\pi}_{t'}$ is from Condition 5.3.3).*

Additionally, for each $t \in [2\colon T]$, let

(iv) $\mathbb{E}_{\pi^*_{2:t}}\left[\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta^*_t)\right\|_2^{2+\alpha}\right] < \infty$ *and* $\mathbb{E}_{\pi^*_{2:t}}\left[\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta^*_t)\dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)})\right\|_2^{2+\alpha}\right] < \infty$ *for all $t' \in [2\colon t]$.*

In Condition 5.5.2 parts (ii) and (iv) above, we assume finite moment conditions that involve the functions $\dot{\pi}_{t'}$ from Condition 5.3.3. This will allow us to control the complexity of the function classes $\left\{ \left[\prod_{t'=2}^{t} \pi_{t'}(\,\cdot\,;\beta_{t'-1}) \right] \varphi_t(\,\cdot\,;\beta_t) \text{ s.t. } \beta_{1:t} \in B_{1:t} \right\}$ for $t \in [2\colon T-1]$; see the Remark below Theorem C.2.2 in Appendix C.2.3 for more details. Note that these assumptions involving $\dot{\pi}_{t'}$ in Condition 5.5.2 are relatively mild and are satisfied if $\mathbb{E}_{\pi^*_{2:t}}\left[\left| \dot{\pi}_t(A_t^{(i)}, S_t^{(i)}) \right|^{4+2\alpha} \right] < \infty$ for all $t \in [2\colon T]$, and $\mathbb{E}_{\pi^*_{2:t}}\left[\left| \dot{\varphi}_t(\mathcal{H}_t^{(i)}) \right|^{4+2\alpha} \right] < \infty$ and $\mathbb{E}_{\pi^*_{2:t}}\left[\left\| \varphi_t(\mathcal{H}_t^{(i)}; \beta_t^*) \right\|_2^{4+2\alpha} \right] < \infty$ for all $t \in [1\colon T-1]$.

Remark 5.5.1 (More General Policy Estimating Functions)**.** *We use Condition 5.5.2 to help ensure a stochastic equicontinuity result holds for the policy parameters. Condition 5.5.2 can be replaced by more general conditions involving bracketing numbers for function classes* $\left\{ \varphi_t(\,\cdot\,;\beta_t) \text{ s.t. } \beta_t \in B_t \right\}$ *for* $t \in [1\colon T-1]$*. See Appendix C.2.3 for the statement of the stochastic equicontinuity result and the more general sufficient conditions.*

Theorem Statements

We have two main theorems. The first, Theorem 5.5.1, shows the consistency of $\hat{\theta}$, i.e., that $\hat{\theta} \xrightarrow{P} \theta^*$. The second, Theorem 5.5.2, shows that $\sqrt{n}\left(\hat{\theta} - \theta^*\right)$ is asymptotically normal with the adaptive sandwich limiting variance from display (5.5.2). Conditions 5.3.1-5.5.2 introduced earlier in Sections 5.3.2 and 5.5.2 are all the assumptions we make on the adaptive sampling algorithm for these two theorems. The remaining assumptions will concern the Z-estimation function ψ (used to define the inferential target θ^* and estimator $\hat{\theta}$). We will require that ψ "plays nicely" with the adaptive sampling algorithm. In other words, there may be choices of ψ that make it incompatible with the adaptive sampling algorithm used to collect the data.

We now state our main theorems. In the conditions for these theorems, we use bracketing numbers to control the complexity of function classes. The bracketing number of a class of real-valued

functions $\mathcal{F}$ is the number of brackets, i.e., pairs functions, of a certain "size" needed to cover $\mathcal{F}$. Following the notation used in Chapter 19 of[99], for any function class $\mathcal{F}$ of real-valued, measurable functions of $\mathcal{H}_T^{(i)}$, we use $N_{[\,]}\big(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)$ to denote the number of brackets of size ε in $L_p(\mathcal{P}_{\pi^*})$ norm needed to cover $\mathcal{F}$; see Appendix C.1.4 for a formal definition of bracketing numbers.

Theorem 5.5.1 (Consistency). *We assume that Conditions 5.3.1-5.3.3 hold for the adaptive sampling algorithm. Then*

$$\hat{\theta} \xrightarrow{P} \theta^*$$

under the following assumptions on the estimator $\hat{\theta}$ and its corresponding estimating function ψ:

(C1) ***Well-Separated Solution:*** *For any $\varepsilon > 0$, there exists some $\eta > 0$ such that*

$$\inf_{\theta \in \mathbb{R}^{d_\theta} \text{ s.t. } \|\theta - \theta^*\|_1 > \varepsilon} \left\| \mathbb{E}_{\pi^*_{2:T}}\big[\psi(\mathcal{H}_T^{(i)}; \theta)\big] \right\|_1 > \eta > 0.$$

(C2) ***Asymptotically Tight:*** *For any $\varepsilon > 0$, there exists some $k < \infty$ such that*

$$\limsup_{n \to \infty} \mathbb{P}\big(\big\|\hat{\theta}\big\|_1 > k\big) \leq \varepsilon.$$

(C3) ***Finite Bracketing Number:*** *Let $\alpha > 0$ be a constant. For any compact subset $K_\theta \subset \mathbb{R}^{d_\theta}$,*

(i) For any $\varepsilon > 0$ and any vector $c \in \mathbb{R}^{d_\theta}$, the bracketing number

$$N_{[\,]}\Big(\varepsilon,\ \big\{c^\top \psi(\,\cdot\,; \theta) \text{ s.t. } \theta \in K_\theta\big\},\ L_{1+\alpha}(\mathcal{P}_{\pi^*})\Big) < \infty. \tag{5.5.19}$$

(ii) There exists a function F_ψ where for any $\theta \in K_\theta$, $\big\|\psi(\mathcal{H}_T^{(i)}; \theta)\big\|_1 \leq F_\psi(\mathcal{H}_T^{(i)})$ a.s. and

$$\mathbb{E}_{\pi^*_{2:T}}\left[\left|F_\psi(\mathcal{H}_T^{(i)})\dot{\pi}_t(A_t^{(i)}, S_t^{(i)})\right|^{1+\alpha}\right] < \infty \tag{5.5.20}$$

for all $t \in [2\colon T]$; the functions $\dot{\pi}_t$ above are from Condition 5.3.3.

Assumption (C1) is used to ensure that there is a well-separated solution for the inferential quantity of interest, θ^*; this also ensures that θ^* is a unique root of the Z-estimation criterion from display (5.3.1). This well-separated condition is commonly used in consistency proofs for Z-estimators.

Next, assumption (C2) is used to ensure that the estimator $\hat{\theta}$ is asymptotically tight, i.e., does not tend towards infinity. In general, this assumption holds when the θ parameter space is bounded or when the estimating function ψ is the derivative of a concave function, e.g., see Theorem 5.14 of[99].

Finally, assumption (C3) restricts the complexity of the function classes

$\left\{c^\top \psi(\,\cdot\,;\theta) \text{ s.t. } \theta \in K_\theta\right\}$ via bracketing numbers. For i.i.d. data, finite bracketing number conditions akin to display (5.5.19) of assumption (C3) are used to show uniform law of large number results[100] Theorem 19.4. However, since the adaptively sampled data is not i.i.d., we use assumption (C3) to prove a martingale version of a uniform law of large numbers (Theorem C.4.2). Specifically, our martingale uniform law of large numbers will concern functions weighted by Radon-Nikodym derivatives. To facilitate use of these Radon-Nikodym weights we control the bracketing complexity of the function class

$$\mathcal{F}_{\Pi c^\top \psi}(B_{1:T-1}, K_\theta) \triangleq \left\{ \left[\prod_{t=2}^{T} \pi_t(\,\cdot\,;\beta_{t-1}) \right] c^\top \psi(\,\cdot\,;\theta) \text{ s.t. } \beta_{1:T-1} \in B_{1:T-1}, \theta \in K_\theta \right\}. \tag{5.5.21}$$

In particular, we use display (5.5.20) of assumption (C3) to help ensure that for any $\varepsilon > 0$ and any $c \in \mathbb{R}^{d_\theta}$, $N_{[\,]}\left(\varepsilon,\ \mathcal{F}_{\Pi c^\top \psi}(B_{1:T-1}, K_\theta),\ L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$; see Lemma C.1.4.

Theorem 5.5.2 (Asymptotic Normality)**.** *We assume that Conditions 5.3.1-5.5.2 hold for the adaptive sampling algorithm. Furthermore, we assume that $\hat{\theta} \xrightarrow{P} \theta^*$ holds (result of Theorem 5.5.1). Then, for $\dot{\Psi}^* \triangleq \frac{\partial}{\partial \theta} \mathbb{E}_{\pi^*_{2:T}}\left[\psi\left(\mathcal{H}_T^{(i)};\theta\right)\right]\big|_{\theta=\theta^*}$ and for Σ^{adapt} as defined in display (5.5.3),*

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{D} \mathcal{N}\left(0,\ [\dot{\Psi}^*]^{-1} \Sigma^{\text{adapt}} [\dot{\Psi}^*]^{-1,\top}\right),$$

under the following additional assumptions:

(N1) ***Invertible "Bread":*** *The mapping* $\theta \mapsto \mathbb{E}_{\pi^*_{2:T}}\big[\psi\big(\mathcal{H}_T^{(i)};\theta\big)\big]$ *is differentiable at* $\theta = \theta^*$. *Moreover, the derivative matrix* $\dot{\Psi}^* \triangleq \frac{\partial}{\partial\theta}\mathbb{E}_{\pi^*_{2:T}}\big[\psi\big(\mathcal{H}_T^{(i)};\theta\big)\big]\big|_{\theta=\theta^*}$ *is invertible.*

(N2) ***Differentiable with Respect to Policy Parameters:***

The mapping $\beta_{1:T-1} \mapsto \mathbb{E}_{\pi_{2:T}(\beta_{1:T-1})}\big[\psi\big(\mathcal{H}_T^{(i)};\theta^*\big)\big]$ *is differentiable at* $\beta_{1:T-1} = \beta^*_{1:T-1}$.

(N3) ***Continuity Condition:*** *The following mapping is continuous at* $\theta = \theta^*$*:*

$$\theta \mapsto \mathbb{E}_{\pi^*_{2:T}}\left[\left\|\psi\big(\mathcal{H}_T^{(i)};\theta\big) - \psi\big(\mathcal{H}_T^{(i)};\theta^*\big)\right\|_2^2\right].$$

(N4) ***Finite Bracketing Integral:*** *Let* $\Theta \subset \mathbb{R}^{d_\theta}$ *be a compact subset whose interior contains* θ^* *and let* $\alpha > 0$ *be a constant.*

(i) For any vector $c \in \mathbb{R}^{d_\theta}$*,*

$$\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon,\ \left\{c^\top\psi(\,\cdot\,;\theta)\ \text{s.t.}\ \theta\in\Theta\right\},\ L_{2+\alpha}(\mathcal{P}_{\pi^*})\right)}\,d\varepsilon < \infty. \tag{5.5.22}$$

(ii) There exists a function F_ψ *such that for all* $\theta \in \Theta$, $\left\|\psi(\mathcal{H}_T^{(i)};\theta)\right\|_1 \le F_\psi(\mathcal{H}_T^{(i)})$ *a.s. and*

$$\mathbb{E}_{\pi^*_{2:T}}\left[\left|F_\psi(\mathcal{H}_T^{(i)})\dot{\pi}_t(A_t^{(i)},S_t^{(i)})\right|^{2+\alpha}\right] < \infty \tag{5.5.23}$$

for all $t \in [2\!:\!T]$*; the functions* $\dot{\pi}_t$ *above are from Condition 5.3.3.*

We now discuss each of the four parts of the assumptions of Theorem 5.5.2 (Asymptotic Normality). Assumption (N1) is used to ensure the "bread" part of the adaptive sandwich variance, $\dot{\Psi}^*$, is invertible. Note that the standard sandwich variance for Z-estimators on i.i.d. data also requires this condition to hold; see the discussion above display (5.5.1).

Assumption (N2) ensures that the estimating function $\mathbb{E}_{\pi_{2:T}(\beta_{1:T-1})}\left[\psi\left(\mathcal{H}_T^{(i)}; \theta^*\right)\right]$ is differentiable with respect to the policy parameters $\beta_{1:T-1}$. This condition is specific to the adaptively sampled data setting. The derivative $\frac{\partial}{\partial \beta_{1:T-1}} \mathbb{E}_{\pi_{2:T}(\beta_{1:T-1})}\left[\psi\left(\mathcal{H}_T^{(i)}; \theta^*\right)\right]\big|_{\beta_{1:T-1}=\beta^*_{1:T-1}}$ is a component of the terms M_t from Σ^{adapt} in the adaptive sandwich variance; see display (5.5.3) for the definition of Σ^{adapt} and see display (5.5.32) for the definition of M_t.

Assumption (N3) is used to ensure that if $\hat{\theta} \xrightarrow{P} \theta^*$, then the $L_2(\mathcal{P}_{\pi^*})$ distance between $\psi\left(\cdot\,; \hat{\theta}\right)$ and $\psi\left(\cdot\,; \theta^*\right)$ converges in probability to zero. This type of condition is classically used in the empirical processes literature to show stochastic equicontinuity results[99] Lemma 19.24.

Finally, display (5.5.22) of assumption (N4) is a finite bracketing integral condition on the function class $\left\{c^\top \psi(\,\cdot\,; \theta) \text{ s.t. } \theta \in \Theta\right\}$ for any $c \in \mathbb{R}^{d_\theta}$. Note that if this function class is Lipschitz, this is sufficient for the finite bracketing integral condition in display (5.5.22) to hold[99] Example 19.7. For i.i.d. data, this kind of bracketing integral condition is used to show functional asymptotic normality results[99] Theorem 19.5. Similarly, we show a functional asymptotic normality result for adaptively sampled data using display (5.5.22); this was discussed in more detail earlier in Section 5.5.1. Our central limit theorem considers functions weighted by particular Radon-Nikodym derivatives. Due to our use of these Radon-Nikodym weights, we will need to control the bracketing integral of the function class $\mathcal{F}_{\Pi c^\top \psi}\left(B_{1:T-1}, \Theta\right)$ as defined earlier in display (5.5.21). In particular, we use display (5.5.23) of assumption (N4) to help ensure that for any $c \in \mathbb{R}^{d_\theta}$, $\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon,\ \mathcal{F}_{\Pi c^\top \psi}(B_{1:T-1}, \Theta),\ L_{2+\alpha}(\mathcal{P}_{\pi^*})\right)} d\varepsilon < \infty$; see Lemma C.1.4.

5.5.3 Proof Sketch for Theorem 5.5.2 (Asymptotic Normality)

As first mentioned below display (5.5.10), by Lemma C.3.1 (Equivalent Formulations for the Adaptive Sandwich Variance), the following result is sufficient for the Theorem and will be the main

result we show in this proof:

$$\sqrt{n}\begin{pmatrix}\hat{\beta}1:T-1-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{pmatrix} \xrightarrow{D} \mathcal{N}\left(0, \begin{bmatrix}\dot{\Phi}^*_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^*\end{bmatrix}^{-1} \Sigma_{1:T} \begin{bmatrix}\dot{\Phi}^*_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^*\end{bmatrix}^{-1,\top}\right). \tag{5.5.24}$$

The derivative terms $\dot{\Psi}^*$, $V_{T,1:T-1}$, and $\dot{\Phi}^*_{1:T-1}$ above exist by assumptions (N1), (N2), and Condition 5.5.1 respectively.

We now state several equalities and discuss why they hold below:

$$\begin{aligned}
&-\sqrt{n}\begin{bmatrix}\hat{\Phi}^{(n)}_{1:T-1}\big(\hat{\beta}1:T-1\big)-\Phi_{1:T-1}\big(\hat{\beta}1:T-1\big)\\ \underbrace{\hat{\Psi}^{(n)}\big(\hat{\beta}1:T-1,\hat{\theta}\big)}_{=o_P(1/\sqrt{n})}-\Psi\big(\hat{\beta}1:T-1,\hat{\theta}\big)\end{bmatrix}\\
&\qquad\underbrace{=}_{(a)} -\sqrt{n}\begin{bmatrix}\Phi_{1:T-1}\big(\beta^*_{1:T-1}\big)-\Phi_{1:T-1}\big(\hat{\beta}1:T-1\big)\\ \underbrace{\Psi\big(\beta^*_{1:T-1},\theta^*\big)}_{=0}-\Psi\big(\hat{\beta}1:T-1,\hat{\theta}\big)\end{bmatrix}+o_P(1)\\
&\underbrace{=}_{(b)}\sqrt{n}\begin{bmatrix}\dot{\Phi}^*_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^*\end{bmatrix}\begin{bmatrix}\hat{\beta}1:T-1-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{bmatrix}\\
&\qquad+\sqrt{n}o_P\left(\left\|\begin{matrix}\hat{\beta}1:T-1-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{matrix}\right\|_2\right)+o_P(1).
\end{aligned}\tag{5.5.25}$$

Equality (a) above holds since $\hat{\Psi}^{(n)}\big(\hat{\beta}1:T-1,\hat{\theta}\big) = o_P(1/\sqrt{n})$ and $\Psi\big(\beta^*_{1:T-1},\theta^*\big) = 0$ by the definitions of $\hat{\theta}$ and θ^* from displays (5.3.2) and (5.3.1) respectively; also since $\hat{\Phi}^{(n)}_{1:T-1}\big(\hat{\beta}1:T-1\big) = o_P(1/\sqrt{n})$ and $\Phi_{1:T-1}\big(\beta^*_{1:T-1}\big) = 0$ by the definitions of $\hat{\beta}t$ and β^*_t from displays (5.3.9) and (5.3.7).

Equality (b) above holds by a Taylor series expansion. Specifically by assumptions (N1) and (N2), the mapping $\left(\beta_{1:T-1}, \theta\right) \mapsto \Psi(\beta_{1:T-1}, \theta)$ is differentiable at $\left(\beta_{1:T-1}, \theta\right) = \left(\beta^*_{1:T-1}, \theta^*\right)$. Also the mapping $\beta_{1:T-1} \mapsto \Phi_{1:T-1}(\beta_{1:T-1})$ is differentiable at $\beta_{1:T-1} = \beta^*_{1:T-1}$ by Condition 5.5.1. As mentioned below display (5.5.12), since $\Phi_{1:T-1}(\beta^*_{1:T-1})$ is not a function of θ, $\frac{\partial}{\partial\theta}\Phi_{1:T-1}(\beta^*_{1:T-1})\big|_{\theta=\theta^*} = 0$.

We now state the next set of results and discuss why they hold below:

$$
\begin{aligned}
&-\sqrt{n}\begin{bmatrix} \hat{\Phi}^{(n)}_{1:T-1}\left(\hat{\beta}1:T-1\right) - \Phi_{1:T-1}\left(\hat{\beta}1:T-1\right) \\ \hat{\Psi}^{(n)}\left(\hat{\beta}1:T-1,\hat{\theta}\right) - \Psi\left(\hat{\beta}1:T-1,\hat{\theta}\right) \end{bmatrix} \\
&\qquad \underbrace{=}_{(c)} -\sqrt{n}\begin{bmatrix} \hat{\Phi}^{(n)}_{1:T-1}\left(\beta^*_{1:T-1}\right) - \Phi_{1:T-1}\left(\beta^*_{1:T-1}\right) \\ \hat{\Psi}^{(n)}\left(\beta^*_{1:T-1},\theta^*\right) - \Psi\left(\beta^*_{1:T-1},\theta^*\right) \end{bmatrix} + o_P(1) \underbrace{\xrightarrow{D}}_{(d)} \mathcal{N}\left(0,\Sigma_{1:T}\right). \quad (5.5.26)
\end{aligned}
$$

Equality (c) above is an stochastic equicontinuity result that holds by applying Lemma C.6.1 (Stochastic Equicontinuity). Lemma C.6.1 uses the fact that $\hat{\theta} \xrightarrow{P} \theta^*$ (by assumption of the Theorem) and $\hat{\beta}1:T-1 \xrightarrow{P} \beta^*_{1:T-1}$ (by Condition 5.3.1). Ensuring that the other assumptions needed to apply Lemma C.6.1 are satisfied is more involved; see the proofs of Theorems C.2.2 and 5.5.2 for more details. The proof of Lemma C.6.1 relies on a functional asymptotic normality result that we prove holds on adaptively sampled data for functions weighted by the Radon-Nikodym derivatives (see Section 5.5.1).

Asymptotic normality result (d) above holds by Theorem C.5.1 (Weighted Martingale Triangular Array Central Limit Theorem). This Theorem proves a martingale central limit theorem for functions weighted by the Radon-Nikodym derivatives on adaptively sampled data.

By consolidating the results from displays (5.5.25) and (5.5.26) above, and applying Slutsky's

theorem we get the following result:

$$\sqrt{n}\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}\begin{bmatrix} \hat{\beta}1:T-1-\beta^*_{1:T-1} \\ \hat{\theta}-\theta^* \end{bmatrix} + \sqrt{n}o_P\left(\left\|\begin{matrix} \hat{\beta}1:T-1-\beta^*_{1:T-1} \\ \hat{\theta}-\theta^* \end{matrix}\right\|_2\right) \xrightarrow{D} \mathcal{N}\left(0,\Sigma_{1:T}\right).$$

Note that $\dot{\Psi}^*$ is invertible by assumption (N1) and $\dot{\Phi}^*_{1:T-1}$ is invertible by Condition 5.5.1 and Lemma C.2.2. By Proposition C.2.1 (Blockwise Inversion of Matrices), this is sufficient for $\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}$ to be invertible. Thus, by the continuous mapping theorem,

$$\sqrt{n}\begin{bmatrix} \hat{\beta}1:T-1-\beta^*_{1:T-1} \\ \hat{\theta}-\theta^* \end{bmatrix} + \sqrt{n}O(1)o_P\left(\left\|\begin{matrix} \hat{\beta}1:T-1-\beta^*_{1:T-1} \\ \hat{\theta}-\theta^* \end{matrix}\right\|_2\right) \xrightarrow{D} \mathcal{N}\left(0,\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1}\Sigma_{1:T}\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1,\top}\right).$$

The above implies that $\sqrt{n}\begin{bmatrix} \hat{\beta}1:T-1-\beta^*_{1:T-1} \\ \hat{\theta}-\theta^* \end{bmatrix} = O_P(1)$. This means that $\sqrt{n}O(1)o_P\left(\left\|\begin{matrix} \hat{\beta}1:T-1-\beta^*_{1:T-1} \\ \hat{\theta}-\theta^* \end{matrix}\right\|_2\right) = o_P(1)$. Thus, display (5.5.24) holds by Slutky's Theorem.

5.5.4 Cases in which the Adaptive and Standard Sandwich Variances

In this section, we now discuss formally the conditions under which the adaptive sandwich variance equals the standard sandwich variance. Specifically, we show that this equivalence holds under a property of estimands θ^*, which we call *policy invariance*.

Definition 5.5.1 (Policy Invariance)**.** *We say that θ^* is policy invariant if*

$$0 = V_{T,1:T-1} \triangleq \frac{\partial}{\partial \beta_{1:T-1}} \mathbb{E}_{\pi(\beta_{1:T-1})} \left[\psi\left(\mathcal{H}_T^{(i)}; \theta\right) \right] \Big|_{\beta_{1:T-1} = \beta^*_{1:T-1}}. \tag{5.5.27}$$

We use also use the notation $\left[V_{T,1}, V_{T,2}, \ldots, V_{T,T-1}\right] \triangleq V_{T,1:T-1} \in \mathbb{R}^{d_\theta \times d_{1:T-1}}$.

Using the Radon-Nikodym derivative weighting from (5.5.5) we can equivalently write $V_{T,t}$ as follows:

$$V_{T,t} \triangleq \frac{\partial}{\partial \beta_t} \mathbb{E}_{\pi^*_{2:T}} \left[W_{t+1}^{(i)}(\beta_t, \beta_t^*) \psi(\mathcal{H}_T^{(i)}; \theta^*) \right] \Big|_{\beta_t = \beta_t^*} \in \mathbb{R}^{d_\theta \times d_t}. \tag{5.5.28}$$

We can interpret $V_{T,t}$ as how the estimating function for θ^* changes with small changes in the limiting policy parameter β_t^*. A particular case in which θ^* is policy invariant was first discussed informally in Remark 5.3.1; specifically when ψ is chosen to be a derivative of the likelihood function of a model for a particular outcome, if that model is correctly specified then the estimand θ^* will not be a projection and will not depend on the target policy parameters $\beta^*_{1:T-1}$.

Recall from display (5.5.3) that the adaptive sandwich variance is $[\dot{\Psi}^*]^{-1} \Sigma^{\text{adapt}} [\dot{\Psi}^*]^{-1}$ where

$$\Sigma^{\text{adapt}} \triangleq \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) + \dot{\Psi}^* \sum_{t=1}^{T-1} M_t \, \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right\}^{\otimes 2} \right]. \tag{5.5.29}$$

We now discuss how the policy invariance property will ensure that the sandwich and adaptive sandwich variances are equivalent.

Recall from the proof sketch from Section 5.5.3 that by Lemma C.3.1 (Equivalent Formulations for the Adaptive Sandwich Variance), $[\dot{\Psi}^*]^{-1}\Sigma^{\text{adapt}}[\dot{\Psi}^*]^{-1}$, the adaptive sandwich variance, equals the lower-right $d_\theta \times d_\theta$ block of limiting variance from display (5.5.10),

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} \Sigma_{1:T} \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1,\top}. \tag{5.5.30}$$

By Proposition C.2.1 (Blockwise Inversion of Matrices),

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} = \begin{bmatrix} \{\dot{\Phi}^*_{1:T-1}\}^{-1} & 0 \\ -\{\dot{\Psi}^*\}^{-1} V_{T,1:T-1} \{\dot{\Phi}^*_{1:T-1}\}^{-1} & \{\dot{\Psi}^*\}^{-1} \end{bmatrix}. \tag{5.5.31}$$

The matrices $M_t \in \mathbb{R}^{d_\theta \times d_t}$ in Σ^{adapt} above are defined as the lower left block of the inverse matrix above in display (5.5.31):

$$\big[M_1, M_2, \ldots, M_{T-1}\big] \triangleq -\{\dot{\Psi}^*\}^{-1} V_{T,1:T-1} \{\dot{\Phi}^*_{1:T-1}\}^{-1} \in \mathbb{R}^{d_\theta \times d_{1:T-1}}. \tag{5.5.32}$$

Above we use $d_{1:T-1} \triangleq \sum_{t=1}^{T-1} d_t$. It is clear from the definition of M_t from display (5.5.32) above that if $V_{T,1:T-1} = 0$ (i.e., the policy invariance property holds), then $\Sigma^{\text{adapt}} = \mathbb{E}_{\pi^*_{2:T}}\big[\psi\big(\mathcal{H}_T^{(i)}; \theta^*\big)^{\otimes 2}\big]$, and the limiting sandwich and adaptive sandwich variances are equivalent.

Note that the asymptotic normality result with the standard sandwich variance for adaptively sampled longitudinal data does not follow from existing results (see Section 5.4 for further discussion of related work). Additionally, when the limiting standard and adaptive sandwich variances are equal, their variance estimators in general will not be equal in small samples (see the formulas for both the sandwich and adaptive sandwich variance estimators in Appendix C.1.1). In general, we advocate for using the adaptive sandwich variance over the standard sandwich variance since it is rare

in digital intervention experiments for inference models to be exactly correctly specified.

5.6 Simulation Results

5.6.1 Data Generation

We consider a binary action setting with $T = 50$ decision times. We use a Boltzmann (or Softmax) exploration type adaptive sampling algorithm [9,20,92]. The state used by the algorithm is the previous time step's reward, i.e., $S_t^{(i)} = \left[1, R_{t-1}^{(i)}\right]$. Specifically, the adaptive sampling algorithm forms action selection probabilities as follows:

$$\pi_t\left(1, S_t^{(i)}; \beta_{t-1}\right) = \text{Clip}_{0.1}\left[\text{expit}\left(\rho \cdot \beta_{t-1,1}^\top S_t^{(i)}\right)\right], \tag{5.6.1}$$

where $\text{Clip}_{0.1}[x] \triangleq \min(\max(x, 0.1), 0.9)$. We can interpret $\beta_{t-1,1}^\top S_t^{(i)}$ above as the adaptive sampling algorithm's working model of a treatment effect. Above the parameter ρ is a positive constant that controls the steepness of the Softmax function; larger values of ρ make the Softmax function steeper. We vary the value of ρ in our experiments. The policy parameter estimators $\hat{\beta}t = \left[\hat{\beta}t, 0, \hat{\beta}t, 1\right]$ are those from the least squares example defined earlier in display (5.3.8).

We generate the rewards as follows:

$$R_t^{(i)} = \kappa_0 + \kappa_1\left[\frac{1}{c_\gamma}\sum_{t'=1}^{t-1}\gamma^{t-1-t'}A_{t'}^{(i)}\right] + \kappa_2 A_t^{(i)} + \varepsilon_t^{(i)}. \tag{5.6.2}$$

Above, the errors $\varepsilon_t^{(i)}$ are generated so that they are correlated over time within a user. We use $\gamma = 0.95$, so $\sum_{t'=1}^{t-1}\gamma^{t-1-t'}A_{t'}^{(i)}$ is a discounted sum of the user's recent "dosage", i.e., the number of times action $A_{t'}^{(i)} = 1$ was previously chosen for that user. We create this dosage variable because the impact of dosage on user receptivity is of great interest in mobile health trials [67,96]. The dosage is

normalized by $c_\gamma \triangleq 1/(1-\gamma)$ to ensure the variable is between $[0, 1]$. In our experiments, we vary the magnitude of the dosage coefficient κ_1. Note that the reward model used by the adaptive sampling algorithm, from display (5.3.8), is incorrectly specified, since it does not take dosage into account. See Appendix C.1.1 for more details.

5.6.2 Data Analysis

For inference we use the following choice of ψ, which corresponds to a least squares criterion:

$$\psi\big(\mathcal{H}_T^{(i)}; \theta\big) = \sum_{t=1}^{T} \big(R_t^{(i)} - \theta_0^\top S_t^{(i)} - \theta_1 A_t^{(i)}\big) \begin{bmatrix} S_t^{(i)} \\ A_t^{(i)} \end{bmatrix}. \tag{5.6.3}$$

Above $\theta_1 \in \mathbb{R}$ parameterizes the marginal treatment effect and $\theta_0 \in \mathbb{R}^2$ parameterize the model of the reward under $A_t^{(i)} = 0$. Note that the data analysis model is misspecified because it does not take into account the dosage variable, which was used to generate rewards as described in display (5.6.2). Thus, the parameters θ_1^* and θ_0^* are projections. When we vary the coefficient for the dosage variable, κ_1, we can see how results are affected by the degree of model misspecification.

We construct 95% confidence intervals for θ_1^*, the projection of the treatment effect. We compare the empirical coverage of confidence intervals constructed using both the standard sandwich and adaptive sandwich variance estimators. We include the formulas for both the sandwich and adaptive sandwich variance estimators in Appendix C.1.1.

5.6.3 Discussion of Results

As seen in Table 5.1, the adaptive sandwich estimator consistently outperforms the standard sandwich variance estimator across all sample sizes and all simulation variants. Moreover, the performance gap increases with the magnitude of the dosage coefficient κ_1. This pattern is expected be-

cause as we discussed in Section 5.5.4, the adaptive and sandwich variances are equivalent when the inference model ψ is correctly specified; increasing the magnitude of κ_1 in the generative model increases the degree of model misspecification for our inference model from display (5.6.3) because it does not include dosage.

Additionally, note that the performance gap between the sandwich and adaptive sandwich variances increases with the algorithm's Softmax steepness parameter ρ. Note that when $\rho = 0$, then there is no adaptive sampling because the action selection probabilities from display (5.6.1) always equal 0.5, i.e., $\pi_t\big(1, S_t^{(i)}; \beta_{t-1}\big) = 0.5$ a.s. As one increases the value of ρ, the steeper the Softmax curve becomes, and the more "adaptive" the algorithm is allowed to be (the algorithm is able to make greater changes in the action selection probabilities). Specifically, when we increase ρ, we expect the norm of the matrices $V_{T,t}$ from display (5.5.28) to grow. Recall that

$$V_{T,t} \triangleq \frac{\partial}{\partial \beta_t} \mathbb{E}_{\pi^*_{2:T}} \left[W_{t+1}^{(i)}(\beta_t, \beta_t^*) \psi(\mathcal{H}_T^{(i)}; \theta^*) \right] \Big|_{\beta_t = \beta_t^*} \in \mathbb{R}^{d_\theta \times d_t}$$

captures how the expectation of the estimating function changes with small changes in the policy parameter β_t. Note that as discussed in Section 5.5.4, when $V_{T,t} = 0$ for all $t \in [1\colon T-1]$, then the limiting sandwich and adaptive sandwich variances are equivalent.

5.7 Discussion

On adaptively sampled data the error of $\hat{\pi}_t^{(n)}$ in estimating the target policy π_t^* impacts what data is collected at the t^{th} decision time, and thus the error of future estimated policies ($\hat{\pi}_{t'}^{(n)}$ for $t' > t$) and the final Z-estimator $\hat{\theta}$. A key conceptual contribution of this work is to provide an approach to represent how the errors in the estimated policies impact the final Z-estimator $\hat{\theta}$. In particular, we show that $\hat{\beta}t - 1$, which parameterizes the estimated policy $\hat{\pi}_t^{(n)}$, can be treated like a plug-in estimator for β_{t-1}^* that was fit on the same dataset used to form $\hat{\theta}$. In other words, even though on adaptively sam-

pled data the estimated policy parameters affect the *data collection*, they can be handled analogously to plug-in estimators for nuisance parameters that are used only in the *data analysis*.

The greatest limitation of this work is that it does not apply to adaptively sampled batch datasets in which the policies used to collect the data are (i) not smooth in their parameters or (ii) allow the amount of exploration to go to zero. As discussed in Section 5.5.2, many common RL algorithms for bandit and Markov decision process settings do not satisfy our smoothness and exploration conditions. However, the studies in digital interventions motivating this work will use adaptive sampling algorithms that do satisfy these conditions.

Future work includes deriving efficient estimators based on adaptively sampled data and designing algorithms that are able to effectively pool over heterogeneous users. Another direction for future work will be to allow the estimating function itself, ψ to include the adaptive sampling action selection probabilities. For example, estimating functions used in off-policy analyses commonly include the action selection probabilities used to collect the data [51,53,94]. More generally, it would be of interest to extend this work to allow the estimating function ψ to include different types of plug-in estimators, e.g., plug-in estimates of the Q-function which are also often used in off-policy analyses [51,53,94]. There are also open questions about how to incorporate estimates formed by high-dimensional machine learning models to potentially increase the efficiency of estimators in these settings.

Table 5.1: Empirical Coverage of Confidence 95% Intervals for Projected Treatment Effect θ_1^*. 2000 Monte Carlo repetitions; standard errors are in parentheses.

Dosage Coeff.	Alg. Steepness	Sample Size	Sandwich	Adaptive Sandwich
$\kappa_1 = 1$	$\rho = 0.5$	$n = 50$	93.65% (0.55)	96.95% (0.39)
		$n = 100$	93.45% (0.55)	97% (0.38)
		$n = 500$	93.3% (0.56)	95.5% (0.46)
	$\rho = 1$	$n = 50$	92.3% (0.6)	96.7% (0.4)
		$n = 100$	90.85% (0.65)	97.3% (0.36)
		$n = 500$	89.85% (0.68)	95.6% (0.46)
	$\rho = 5$	$n = 50$	75.8% (0.96)	95.4% (0.47)
		$n = 100$	77.6% (0.93)	96.5% (0.41)
		$n = 500$	73.05% (0.99)	95.7% (0.45)
$\kappa_1 = 5$	$\rho = 0.5$	$n = 50$	86.7% (0.76)	95.15% (0.48)
		$n = 100$	87.6% (0.74)	95.85% (0.45)
		$n = 500$	85.8% (0.78)	94.65% (0.5)
	$\rho = 1$	$n = 50$	83.5% (0.83)	96.2% (0.43)
		$n = 100$	83.65% (0.83)	95.65% (0.46)
		$n = 500$	80.0% (0.89)	95.35% (0.47)
	$\rho = 5$	$n = 50$	54.85% (1.1)	90.9% (0.64)
		$n = 100$	52.85% (1.1)	94.35% (0.52)
		$n = 500$	45.9% (1.1)	95.25% (0.48)

6

Conclusion

In this thesis, we have (i) developed a deeper understanding of the reasons behind the failure of classical statistical methods on adaptively collected data and (ii) developed a variety of methods for statistical inference on adaptively collected data for different environments. These methods facilitate the use of RL and other adaptive algorithms for experiments in which post-study causal inference is critical, e.g., in digital health, online education, and public policy applications. However, there are still a variety of open questions in this area. We describe some of these below.

6.1 Open Questions

6.1.1 Accommodating Alternative Forms of Pooling

In the section on Inference after Adaptive Sampling for Longitudinal Data, we consider a particular class of pooling RL algorithms. In practice there may be other types of pooling RL algorithms which we may want to run. For example, ones that partially pool, e.g., by pooling the data of more similar users more and pooling the data of more dissimilar users less. Additionally, another form of pooling is limited resource allocation. In this setting, how likely a given user is allocated a resource depends on the outcomes of other users.[73] How to construct valid confidence intervals for treatment effects under these types of pooling algorithms is unclear.

6.1.2 Post-Selection Inference on Adaptively Collected Data

Often adaptive experiments that are run involve experimenting with a large number of treatments and a goal after the study is over is to estimate the treatment effect between the estimated best performing treatment and a control treatment. Besides the adaptively collected data issue, this problem is non-standard because the data analyzer uses estimates of the treatments to select which is the best performing treatment, before constructing a confidence interval for it. For accurate inference, one must account for the fact that the treatment used one is forming the construct confidence interval form was selected by looking at estimates formed by the same dataset.[8,47]

6.1.3 Multi-Objective Reinforcement Learning

Broadly, when designing studies using RL algorithms we are often interested in optimizing (1) within-study personalization (or maximizing rewards), which can be achieved using an RL algorithm, and (2) after-study analyses, which typically involve inferring treatment effects. Multiple

works have noted a trade-off between the objectives of maximizing rewards and accurately inferring treatment effects (e.g. minimizing width of a confidence interval).[44,90,32,24] There are other objectives of interest like identifying the best treatment with high probability. There are open questions as to (i) the relationship between these multiple objects, e.g., fundamental trade-offs or equivalencies, and (ii) how to tightly trade-off these objectives. Furthermore, there are currently few works looking at efficiency theory estimation for the adaptively collected data.

A

Inference for Batched Bandits

A.1 Simulation Details

A.1.1 W-Decorrelated Estimator

For the W-decorrelated estimator[26], for a batch size of n and for T batches, we set λ to be the $\frac{1}{nT}$ quantile of $\lambda_{\min}(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})/\log(nT)$, where $\lambda_{\min}(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})$ denotes the minimum eigenvalue of $\underline{\mathbf{X}}^\top \underline{\mathbf{X}}$. This procedure of choosing λ is motivated by the conditions of Theorem 4 of[26] and follows the methods used by[26] in their simulation experiments. We had to adjust the original procedure for choosing λ used by[26] (who set λ to the 0.15 quantile of $\lambda_{\min}(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})$), because they only evaluated the W-decorrelated method for when the total number of samples was $nT = 1000$ and valid values of λ changes with the sample size.

A.1.2 AW-AIPW Estimator

Since the AW-AIPW test statistic for the treatment effect is not explicitly written in the original paper[39], we now write the formulas for the AW-AIPW estimator of the treatment effect: $\hat{\Delta}^{\text{AW-AIPW}} \triangleq \hat{\beta}_1^{\text{AW-AIPW}} - \hat{\beta}_0^{\text{AW-AIPW}}$. We use the variance stabilizing weights, equal to the square root of the sampling probabilities, $\sqrt{\pi_t^{(n)}}$ and $\sqrt{1-\pi_t^{(n)}}$. Below, $N_{t,1} = \sum_{i=1}^n A_{t,i}$ and $N_{t,0} = \sum_{i=1}^n (1 - A_{t,i})$.

$$Y_{t,1} \triangleq \frac{A_{t,i}}{\pi_t^{(n)}} R_{t,i} + \left(1 - \frac{A_{t,i}^{(n)}}{\pi_t^{(n)}}\right) \frac{\sum_{t'=1}^{t-1} \sum_{i=1}^n A_{t,i}^{(n)} R_{t,i}}{\sum_{t'=1}^{t-1} N_{t',1}}$$

$$Y_{t,0} \triangleq \frac{1 - A_{t,i}^{(n)}}{1 - \pi_t^{(n)}} R_{t,i} + \left(1 - \frac{1 - A_{t,i}}{1 - \pi_t^{(n)}}\right) \frac{\sum_{t'=1}^{t-1} \sum_{i=1}^n (1 - A_{t,i}) R_{t,i}}{\sum_{t'=1}^{t-1} N_{t',0}}$$

$$\hat{\beta}_1^{\text{AW-AIPW}} \triangleq \frac{\sum_{t=1}^T \sum_{i=1}^n \sqrt{\pi_t^{(n)}} Y_{t,1}}{\sum_{t=1}^T \sum_{i=1}^n \sqrt{\pi_t^{(n)}}} \quad \text{and} \quad \hat{\beta}_0^{\text{AW-AIPW}} \triangleq \frac{\sum_{t=1}^T \sum_{i=1}^n \sqrt{1 - \pi_t^{(n)}} Y_{t,0}}{\sum_{t=1}^T \sum_{i=1}^n \sqrt{1 - \pi_t^{(n)}}}$$

The variance estimator for $\hat{\Delta}^{\text{AW-AIPW}}$ is $\hat{V}_0 + \hat{V}_1 + 2\hat{C}_{0,1}$ where

$$\hat{V}_1 \triangleq \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}\pi_t^{(n)}(Y_{t,1} - \hat{\beta}_1^{\text{AW-AIPW}})^2}{\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\sqrt{\pi_t^{(n)}}\right)^2} \quad \text{and} \quad \hat{V}_0 \triangleq \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(1-\pi_t^{(n)})(Y_{t,0} - \hat{\beta}_0^{\text{AW-AIPW}})^2}{\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\sqrt{1-\pi_t^{(n)}}\right)^2}$$

$$\hat{C}_{0,1} \triangleq -\frac{\sum_{t=1}^{T}\sum_{i=1}^{n}\sqrt{\pi_t^{(n)}(1-\pi_t^{(n)})}(Y_{t,1} - \hat{\beta}_1^{\text{AW-AIPW}})(Y_{t,0} - \hat{\beta}_0^{\text{AW-AIPW}})}{\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\sqrt{\pi_t^{(n)}}\right)\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\sqrt{1-\pi_t^{(n)}}\right)}$$

A.1.3 Self-Normalized Martingale Bound

By the self-normalized martingale bound of[2], specifically Theorem 1 and Lemma 6, we have that in the two arm bandit setting,

$$\mathbb{P}\left(\forall T, n \geq 1, \left|\hat{\beta}_1^{\text{OLS}} - \beta_1\right| \leq c_{1,T} \text{ and } \left|\hat{\beta}_0^{\text{OLS}} - \beta_0\right| \leq c_{0,T}\right) \geq 1 - \delta$$

where

$$c_{a,T} = \sqrt{\sigma^2 \frac{1 + \sum_{t=1}^{T} N_{t,a}}{\left(\sum_{t=1}^{T} N_{t,a}\right)^2}\left(1 + 2\log\left(\frac{2\sqrt{1 + \sum_{t=1}^{T} N_{t,a}}}{\delta}\right)\right)}$$

We estimate σ^2 using the procedure stated below for the OLS estimator. We reject the null hypothesis that $\Delta = 0$ whenever either the confidence bounds for the two arms are non-overlapping. Specifically when

$$\hat{\beta}_1^{\text{OLS}} + c_{1,T} \leq \hat{\beta}_0^{\text{OLS}} - c_{0,T} \quad \text{or} \quad \hat{\beta}_0^{\text{OLS}} + c_{0,T} \leq \hat{\beta}_1^{\text{OLS}} - c_{1,T}$$

A.1.4 Estimating Noise Variance

OLS Estimator Given the OLS estimators for the means of each arm, $\hat{\beta}_1^{\text{OLS}}, \hat{\beta}_0^{\text{OLS}}$, we estimate the noise variance σ^2 as follows:

$$\hat{\sigma}^2 \triangleq \frac{1}{nT-2} \sum_{t=1}^{T} \sum_{i=1}^{n} \left(R_{t,i} - A_{t,i} \hat{\beta}_1^{\text{OLS}} - (1 - A_{t,i}) \hat{\beta}_0^{\text{OLS}} \right)^2.$$

We use a degrees of freedom bias correction by normalizing by $nT - 2$ rather than nT. Since the W-decorrelated estimator is a modified version of the OLS estimator, we also use this same noise variance estimator for the W-decorrelated estimator; we found that this worked well in practice, in terms of Type-1 error control.

Batched OLS Given the Batched OLS estimators for the means of each arm for each batch, $\hat{\beta}_{t,1}^{\text{BOLS}}, \hat{\beta}_{t,0}^{\text{BOLS}}$, we estimate the noise variance for each batch σ_t^2 as follows:

$$\hat{\sigma}_t^2 \triangleq \frac{1}{n-2} \sum_{i=1}^{n} \left(R_{t,i} - A_{t,i} \hat{\beta}_{t,1}^{\text{BOLS}} - (1 - A_{t,i}) \hat{\beta}_{t,0}^{\text{BOLS}} \right)^2.$$

Again, we use a degrees of freedom bias correction by normalizing by $n - 2$ rather than n. We prove the consistency of $\hat{\sigma}_t^2$ (meaning $\hat{\sigma}_t^2 \xrightarrow{P} \sigma^2$) in Corollary A.4.2. Using BOLS to test $H_0 : \Delta = a$ vs. $H_1 : \Delta \neq a$, we use the following test statistic:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{\frac{N_{t,0} N_{t,1}}{n \hat{\sigma}_t^2}} (\hat{\Delta}_t^{\text{BOLS}} - a).$$

Above, $N_{t,1} = \sum_{i=1}^{n} A_{t,i}$ and $N_{t,0} = \sum_{i=1}^{n}(1 - A_{t,i})$. For this test statistic, we use cutoffs based on the Student-t distribution, i.e., for $Y_t \overset{i.i.d.}{\sim} t_{n-2}$ we use a cutoff $c_{\alpha/2}$ such that

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t\right| > c_{\alpha/2}\right) = \alpha.$$

We found $c_{\alpha/2}$ by simulating draws from the Student-t distribution.

A.1.5 Non-Stationary Treatment Effect

When we believe that the margin itself varies from batch to batch, we are able to construct a confidence region that contains the true margin Δ_t for each batch simultaneously with probability $1 - \alpha$.

Corollary A.1.1 (Confidence band for margin for non-stationary bandits)**.** *Assume the same conditions as Theorem 3.5.1. Let z_x be x^{th} quantile of the standard Normal distribution, i.e., for $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(Z < z_\alpha) = \alpha$. For each $t \in [1\colon T]$, we define the interval*

$$\boldsymbol{L}_t = \hat{\Delta}_t^{\text{OLS}} \pm z_{1-\frac{\alpha}{2T}}\sqrt{\frac{\sigma^2 n}{N_{t,0}N_{t,1}}}.$$

$\lim_{n\to\infty} \mathbb{P}\big(\forall t \in [1:T], \Delta_t \in \boldsymbol{L}_t\big) \geq 1 - \alpha$. *Above, $N_{t,1} = \sum_{i=1}^{n} A_{t,i}$ and $N_{t,0} = \sum_{i=1}^{n}(1 - A_{t,i})$.*

Proof: Note that by Corollary A.4.1,

$$\mathbb{P}\big(\text{exists some } t \in [1:T] \text{ s.t. } \Delta_t \notin \mathbf{L}_t\big) \leq \sum_{t=1}^{T} \mathbb{P}\big(\Delta_t \notin \mathbf{L}_t\big) \to \sum_{t=1}^{T} \frac{\alpha}{T} = \alpha$$

where the limit is as $n \to \infty$. Since

$$\mathbb{P}\big(\forall t \in [1:T], \Delta_t \in \mathbf{L}_t\big) = 1 - \mathbb{P}\big(\text{exists some } t \in [1:T] \text{ s.t. } \Delta_t \notin \mathbf{L}_t\big)$$

Thus,

$$\lim_{n\to\infty} \mathbb{P}\big(\forall t \in [1:T], \Delta_t \in \mathbf{L}_t\big) \geq 1-\alpha \qquad \square$$

We can also test the null hypothesis of no margin against the alternative that at least one batch has non-zero margin, i.e., $H_0: \forall t \in [1:T], \Delta_t = 0$ vs. $H_1: \exists t \in [1:T]\, s.t.\, \Delta_t \neq 0$. Note that the global null stated above is of great interest in the mobile health literature[60,68]. Specifically we use the following test statistic:

$$\sum_{t=1}^{T} \frac{N_{t,0}N_{t,1}}{\sigma^2 n}(\hat{\Delta}_t^{\text{OLS}} - 0)^2,$$

which by Theorem 3.5.1 converges in distribution to a chi-squared distribution with T degrees of freedom under the null $\Delta_t = 0$ for all t.

To account for estimating noise variance σ^2, in our simulations for this test statistic, we use cut-offs based on the Student-t distribution, i.e., for $Y_t \overset{i.i.d.}{\sim} t_{n-2}$ we use a cutoff $c_{\alpha/2}$ such that

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} Y_t^2 > c_\alpha\right) = \alpha.$$

We found c_α by simulating draws from the Student-t distribution.

In the plots below we call the test statistic in (A.1.5) "BOLS Non-Stationary Treatment Effect" (BOLS NSTE). BOLS NSTE performs poorly in terms of power compared to other test statistics in the stationary setting; however, *in the non-stationary setting, BOLS NSTE significantly outperforms all other test statistics, which tend to have low power when the average treatment effect is close to zero.* Note that the W-decorrelated estimator performs well in the left plot of Figure A.3; this is because as we show in Appendix A.6, the W-decorrelated estimator upweights samples from the earlier batches in the study. So when the treatment effect is large in the beginning of the study, the W-decorrelated estimator has high power and when the treatment effect is small or zero in the beginning of the study, the W-decorrelated estimator has low power.

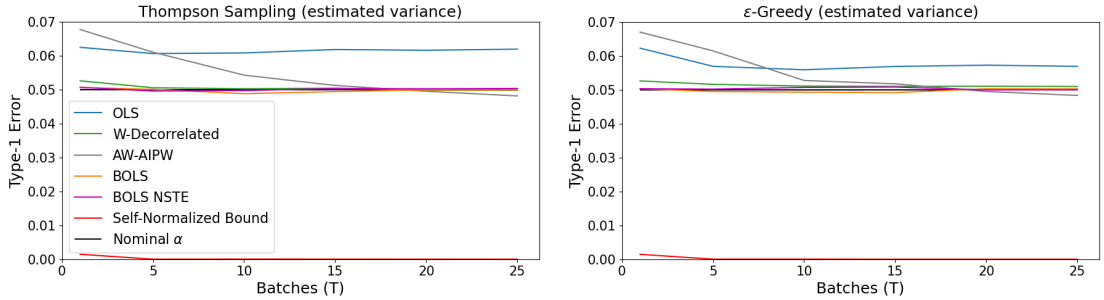

Figure A.1: Stationary Setting: Type-1 error for a two-sided test of $H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0$ ($\alpha = 0.05$). We set $\beta_1 = \beta_0 = 0, n = 25$, and a clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. We use 100k Monte Carlo simulations and standard errors are < 0.001.

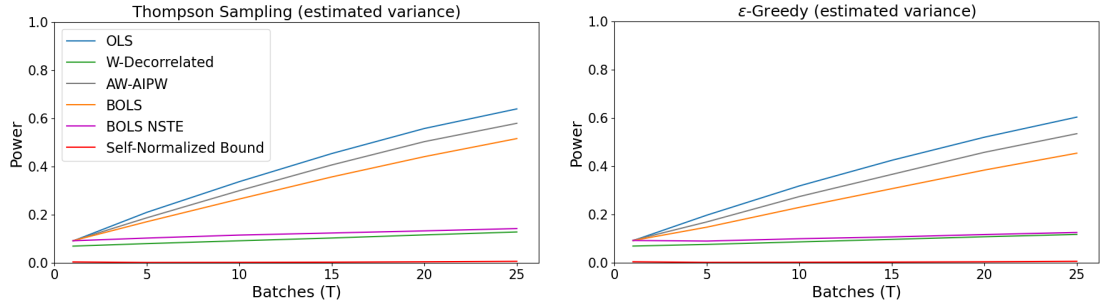

Figure A.2: Stationary Setting: Power for a two-sided test of $H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0$ ($\alpha = 0.05$). We set $\beta_1 = 0$, $\beta_0 = 0.25, n = 25$, and a clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. We use 100k Monte Carlo simulations and standard errors are < 0.002. We account for Type-1 error inflation as described in Section 3.6.

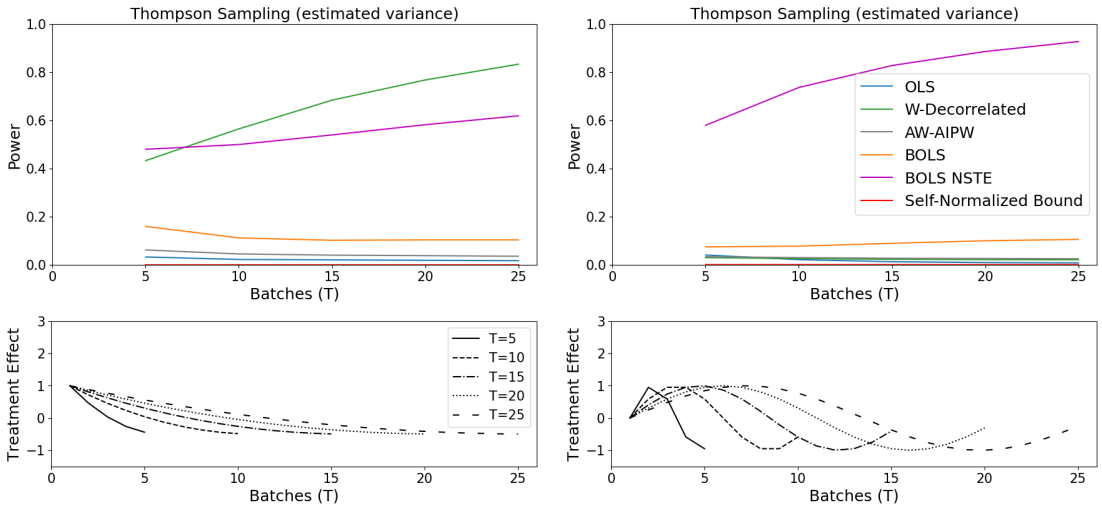

Figure A.3: Nonstationary setting: The two upper plots display the power of estimators for a two-sided test of $H_0 : \forall t \in [1 : T], \beta_{t,1} - \beta_{t,0} = 0$ vs. $H_1 : \exists t \in [1 : T], \beta_{t,1} - \beta_{t,0} \neq 0$ ($\alpha = 0.05$). The two lower plots display two treatment effect trends; the left plot considers a decreasing trend (quadratic function) and the right plot considers a oscillating trend ($\sin$ function). We set $n = 25$, and a clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. We use 100k Monte Carlo simulations and standard errors are < 0.002.

A.2 Asymptotic Normality of the OLS Estimator

Condition A.2.1 (Weak moments). *$\forall t, n, i, \mathbb{E}[\varepsilon_{t,i}^2 | \mathcal{G}_{t-1}^{(n)}] = \sigma^2$ and for all $\forall t, n, i, \mathbb{E}[\phi(\varepsilon_{t,i}^2) | \mathcal{G}_{t-1}^{(n)}] < M < \infty$ a.s. for some function ϕ where $\lim_{x\to\infty} \frac{\phi(x)}{x} \to \infty$.*

Condition A.2.2 (Stability). *There exists a sequence of nonrandom positive-definite symmetric matrices, $\underline{\boldsymbol{V}}_n$, such that*

(a) $\underline{\boldsymbol{V}}_n^{-1} \big(\sum_{t=1}^{T} \sum_{i=1}^{n} \boldsymbol{X}_{t,i} \boldsymbol{X}_{t,i}^\top \big)^{\frac{1}{2}} = \underline{\boldsymbol{V}}_n^{-1} \big(\underline{\boldsymbol{X}}^\top \underline{\boldsymbol{X}} \big)^{\frac{1}{2}} \xrightarrow{P} \underline{\boldsymbol{I}}_p$

(b) $\max_{i \in [1:\, n], t \in [1:\, T]} \| \underline{\boldsymbol{V}}_n^{-1} \boldsymbol{X}_{t,i} \|_2 \xrightarrow{P} 0$

Theorem A.2.1 (Triangular array version of Lai & Wei (1982), Theorem 3). *Let $\boldsymbol{X}_{t,i} \in \mathbb{R}^p$ be non-anticipating with respect to filtration $\{\mathcal{G}_t^{(n)}\}_{t=1}^T$, so $\boldsymbol{X}_{t,i}$ is $\mathcal{G}_{t-1}^{(n)}$ measurable. We assume the following conditional mean model for rewards:*

$$\mathbb{E}\big[R_{t,i} \big| \mathcal{G}_{t-1}^{(n)} \big] = \boldsymbol{X}_{t,i}^\top \beta.$$

We define $\varepsilon_{t,i} \triangleq R_{t,i} - \boldsymbol{X}_{t,i}^\top \beta$. Note that $\{\varepsilon_{t,i}\}_{i=1,t=1}^{i=n,t=T}$ is a martingale difference array with respect to filtration $\{\mathcal{G}_t^{(n)}\}_{t=1}^T$.

Assuming Conditions A.2.1 and A.2.2, as $n \to \infty$,

$$(\underline{\boldsymbol{X}}^\top \underline{\boldsymbol{X}})^{1/2} (\hat{\beta}^{\mathrm{OLS}} - \beta) \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_p)$$

Note, in the body of the paper we state that this theorem holds in the two-arm bandit case assuming Conditions 3.4.2 and 3.4.1. Note that Condition 3.4.1 is sufficient for Condition A.2.1 and Condition 3.4.2 is sufficient for Condition A.2.2 in the two-arm bandit case.

Proof:

$$\hat{\beta}^{\mathrm{OLS}} = ((\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^{(n),\top} \mathbf{R}^{(n)} = (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top (\underline{\mathbf{X}} \beta + \varepsilon)$$

$$\hat{\beta}^{\text{OLS}} - \beta = (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \varepsilon = \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{t,i} \mathbf{X}_{t,i}^\top \right)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{t,i} \varepsilon_{t,i}$$

It is sufficient to show that as $n \to \infty$:

$$(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1/2} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{t,i} \varepsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_p)$$

By Slutsky's Theorem and Condition A.2.2 (a), it is also sufficient to show that as $n \to \infty$,

$$\underline{\mathbf{V}}_n^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{t,i} \varepsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_p)$$

By Cramer-Wold device, it is sufficient to show multivariate normality if for any fixed $\mathbf{c} \in \mathbb{R}^p$ s.t. $\|\mathbf{c}\|_2 = 1$, as $n \to \infty$,

$$\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{t,i} \varepsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

We will prove this central limit theorem by using a triangular array martingale central limit theorem, specifically Theorem 2.2 of[29]. We will do this by letting $Y_{t,i} = \mathbf{c}^T \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \varepsilon_{t,i}$. The theorem states that as $n \to \infty$, $\sum_{t=1}^{T} \sum_{i=1}^{n} Y_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ if the following conditions hold as $n \to \infty$:

(a) $\sum_{t=1}^{T} \sum_{i=1}^{n} E[Y_{t,i} | \mathcal{G}_{t-1}^{(n)}] \xrightarrow{P} 0$

(b) $\sum_{t=1}^{T} \sum_{i=1}^{n} E[Y_{t,i}^2 | \mathcal{G}_{t-1}^{(n)}] \xrightarrow{P} \sigma^2$

(c) $\forall \delta > 0, \sum_{t=1}^{T} \sum_{i=1}^{n} E\left[Y_{t,i}^2 \mathbb{I}_{(|Y_{t,i}| > \delta)} \middle| \mathcal{G}_{t-1}^{(n)} \right] \xrightarrow{P} 0$

Useful Properties Note that by Cauchy-Schwartz and Condition A.2.2 (b), as $n \to \infty$,

$$\max_{i \in [1:\, n], t \in [1:\, T]} \left| \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \right| \leq \max_{i \in [1:\, n], t \in [1:\, T]} \|\mathbf{c}\|_2 \|\underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i}\|_2 \xrightarrow{P} 0$$

By continuous mapping theorem and since the square function on non-negative inputs is order preserving, as $n \to \infty$,

$$\left(\max_{i \in [1:\, n], t \in [1:\, T]} \left| \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \right| \right)^2 = \max_{i \in [1:\, n], t \in [1:\, T]} \left(\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \right)^2 \xrightarrow{P} 0 \tag{A.2.1}$$

By Condition A.2.2 (a) and continuous mapping theorem, $\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} (\mathbf{X}_{t,i}^\top \mathbf{X}_{t,i})^{1/2} \xrightarrow{P} \mathbf{c}^\top$, so

$$\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} (\mathbf{X}_{t,i}^\top \mathbf{X}_{t,i})^{1/2} (\mathbf{X}_{t,i}^\top \mathbf{X}_{t,i})^{1/2} \underline{\mathbf{V}}_n^{-1} \mathbf{c} \xrightarrow{P} \mathbf{c}^\top \mathbf{c} = 1$$

Thus,

$$\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{t,i} \mathbf{X}_{t,i}^\top \right) \underline{\mathbf{V}}_n^{-1} \mathbf{c} = \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \mathbf{X}_{t,i}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{c} \xrightarrow{P} 1$$

Since $\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i}$ is a scalar, as $n \to \infty$,

$$\sum_{t=1}^{T} \sum_{i=1}^{n} (\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i})^2 \xrightarrow{P} 1 \tag{A.2.2}$$

Condition (a): Martingale

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}[\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \varepsilon_{t,i} | \mathcal{G}_{t-1}^{(n)}] = \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \mathbb{E}[\varepsilon_{t,i} | \mathcal{G}_{t-1}^{(n)}] = 0$$

Condition (b): Conditional Variance

$$\begin{aligned} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}[(\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i})^2 \varepsilon_{t,i}^2 | \mathcal{G}_{t-1}^{(n)}] &= \sum_{t=1}^{T} \sum_{i=1}^{n} (\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i})^2 \mathbb{E}[\varepsilon_{t,i}^2 | \mathcal{G}_{t-1}^{(n)}] \\ &= \sigma^2 \sum_{t=1}^{T} \sum_{i=1}^{n} (\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i})^2 \xrightarrow{P} \sigma^2 \end{aligned}$$

where the last equality holds by Condition A.2.1 and the limit holds by (A.2.2) as $n \to \infty$.

CONDITION (C): LINDEBERG CONDITION Let $\delta > 0$. We want to show that as $n \to \infty$,

$$\sum_{t=1}^{T}\sum_{i=1}^{n} Z_{t,i}^2 \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(Z_{t,i}^2\varepsilon_{t,i}^2 > \delta^2)} \middle| \mathcal{G}_{t-1}^{(n)}\right] \xrightarrow{P} 0$$

where above, we define $Z_{t,i}^{(n)} \triangleq \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1}\mathbf{X}_{t,i}$. By Condition A.2.1, we have that for all $n \geq 1$,

$$\max_{t\in[1:\,T],i\in[1:\,n]} \mathbb{E}[\phi(\varepsilon_{t,i}^2)|\mathcal{G}_{t-1}^{(n)}] < M$$

Since we assume that $\lim_{x\to\infty} \frac{\phi(x)}{x} = \infty$, for all $m \geq 1$, there exists a b_m s.t. $\phi(x) \geq mMx$ for all $x \geq b_m$. So, for all n, t, i,

$$M \geq \mathbb{E}[\phi(\varepsilon_{t,i}^2)|\mathcal{G}_{t-1}^{(n)}] \geq \mathbb{E}[\phi(\varepsilon_{t,i}^2)\mathbb{I}_{(\varepsilon_{t,i}^2\geq b_m)}|\mathcal{G}_{t-1}^{(n)}] \geq mM\mathbb{E}[\varepsilon_{t,i}^2\mathbb{I}_{(\varepsilon_{t,i}^2\geq b_m)}|\mathcal{G}_{t-1}^{(n)}]$$

Thus,

$$\max_{t\in[1:\,T],i\in[1:\,n]} \mathbb{E}[\varepsilon_{t,i}^2\mathbb{I}_{(\varepsilon_{t,i}^2\geq b_m)}|\mathcal{G}_{t-1}^{(n)}] \leq \frac{1}{m}$$

So we have that

$$\sum_{t=1}^{T}\sum_{i=1}^{n} Z_{t,i}^2 \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(Z_{t,i}^2\varepsilon_{t,i}^2 > \delta^2)} \middle| \mathcal{G}_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n} Z_{t,i}^2 \left(\mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(Z_{t,i}^2\varepsilon_{t,i}^2 > \delta^2)} \middle| \mathcal{G}_{t-1}^{(n)}\right]\mathbb{I}_{(Z_{t,i}^2 \leq \delta^2/b_m)} + \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(Z_{t,i}^2\varepsilon_{t,i}^2 > \delta^2)} \middle| \mathcal{G}_{t-1}^{(n)}\right]\mathbb{I}_{(Z_{t,i}^2 > \delta^2/b_m)}\right)$$

$$\leq \sum_{t=1}^{T}\sum_{i=1}^{n} Z_{t,i}^2 \left(\mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(\varepsilon_{t,i}^2 > b_m)} \middle| \mathcal{G}_{t-1}^{(n)}\right] + \sigma^2 \mathbb{I}_{(Z_{t,i}^2 > \delta^2/b_m)}\right)$$

$$\leq \left(\frac{1}{m} + \sigma^2 \mathbb{I}_{(\max_{t'\in[1:\,T],j\in[1:\,n]} Z_{t',j}^2 > \delta^2/b_m)}\right) \sum_{t=1}^{T}\sum_{i=1}^{n} Z_{t,i}^2$$

By Slutsky's Theorem and (A.2.2), it is sufficient to show that as $n \to \infty$,

$$\frac{1}{m} + \sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} \xrightarrow{P} 0$$

For any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{m} + \sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} > \varepsilon\right) \leq \mathbb{I}_{(\frac{1}{m} > \frac{\varepsilon}{2})} + \mathbb{P}\left(\sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} > \frac{\varepsilon}{2}\right)$$

We can choose m such that $\frac{1}{m} \leq \frac{\varepsilon}{2}$, so $\mathbb{P}\left(\frac{1}{m} > \frac{\varepsilon}{2}\right) = 0$. For the second term (note that m is now fixed),

$$\mathbb{P}\left(\sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} > \frac{\varepsilon}{2}\right) \leq \mathbb{P}\left(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m\right) \to 0$$

where the last limit holds by (A.2.1) as $n \to \infty$. □

A.2.1 Corollary 3.4.1 (Sufficient conditions for Theorem A.2.1)

Under Conditions 3.4.1 and 3.4.3, when ***the treatment effect is non-zero*** *data collected in batches using ε-greedy, Thompson Sampling, or UCB with a fixed clipping constraint (see Definition 3.3.1) will satisfy Theorem A.2.1 conditions.*

Proof: The only condition of Theorem A.2.1 that needs to verified is Condition 3.4.2. To satisfy Condition 3.4.2, it is sufficient to show that for any given Δ, for some constant $c \in (0, T)$,

$$\frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} A_{t,i} = \frac{1}{n} \sum_{t=1}^{T} N_{t,1} \xrightarrow{P} c.$$

ε-greedy We assume without loss of generality that $\Delta > 0$ and $\pi_1^{(n)} = \frac{1}{2}$. Recall that for ε-greedy, for $a \in [2\colon T]$,

$$\pi_a^{(n)} = \begin{cases} 1 - \frac{\varepsilon}{2} & \text{if } \frac{\sum_{t=1}^{a}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sum_{t'=1}^{a} N_{t',1}} > \frac{\sum_{t=1}^{a}\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{\sum_{t'=1}^{a} N_{t',0}} \\ \frac{\varepsilon}{2} & \text{otherwise} \end{cases}$$

Thus to show that $\pi_a^{(n)} \xrightarrow{P} 1 - \frac{\varepsilon}{2}$ for all $a \in [2\colon T]$, it is sufficient to show that

$$\mathbb{P}\left(\frac{\sum_{t=1}^{a}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sum_{t'=1}^{a} N_{t',1}} > \frac{\sum_{t=1}^{a}\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{\sum_{t'=1}^{a} N_{t',0}} \right) \to 1 \tag{A.2.3}$$

To show (A.2.3), it is equivalent to show that

$$\mathbb{P}\left(\Delta > \frac{\sum_{t=1}^{a}\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sum_{t'=1}^{a} N_{t',0}} - \frac{\sum_{t=1}^{a}\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sum_{t'=1}^{a} N_{t',1}} \right) \to 1 \tag{A.2.4}$$

To show (A.2.4), it is sufficient to show that

$$\frac{\sum_{t=1}^{a}\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sum_{t'=1}^{a} N_{t',0}} - \frac{\sum_{t=1}^{a}\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sum_{t'=1}^{a} N_{t',1}} \xrightarrow{P} 0. \tag{A.2.5}$$

To show (A.2.5), it is equivalent to show that

$$\sum_{t=1}^{a} \frac{\sqrt{N_{t,0}}}{\sum_{t'=1}^{a} N_{t',0}} \frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sqrt{N_{t,0}}} - \sum_{t=1}^{a} \frac{\sqrt{N_{t,1}}}{\sum_{t'=1}^{a} N_{t',1}} \frac{\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{t,1}}} \xrightarrow{P} 0. \tag{A.2.6}$$

By Lemma A.4.1, for all $t \in [1\colon T]$,

$$\frac{N_{t,1}}{\pi_t^{(n)} n} \xrightarrow{P} 1$$

Thus by Slutsky's Theorem, to show (A.2.6), it is sufficient to show that

$$\sum_{t=1}^{a} \frac{\sqrt{n(1-\pi_t^{(n)})}}{n\sum_{t'=1}^{a}(1-\pi_{t'}^{(n)})} \frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sqrt{N_{t,0}}} - \sum_{t=1}^{a} \frac{\sqrt{n\pi_t^{(n)}}}{n\sum_{t'=1}^{a}\pi_{t'}^{(n)}} \frac{\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{t,1}}} \xrightarrow{P} 0. \tag{A.2.7}$$

Since $\pi_t^{(n)} \in [\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{2}]$ for all t, n, the left hand side of (A.2.7) equals the following:

$$\sum_{t=1}^{a} o_p(1) \frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sqrt{N_{t,0}}} - \sum_{t=1}^{a} o_p(1) \frac{\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{t,1}}} \xrightarrow{P} 0.$$

The above limit holds because by Thereom 3.5.1, we have that

$$\left(\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}, ..., \frac{\sum_{i=1}^{n}A_{T,i}\varepsilon_{T,i}}{\sqrt{N_{T,1}}}, \frac{\sum_{i=1}^{n}(1-A_{T,i})\varepsilon_{T,i}}{\sqrt{N_{T,0}}} \right) \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{I}_{2T}). \tag{A.2.8}$$

Thus, by Slutsky's Theorem and Lemma A.4.1, we have that

$$\frac{1}{n}\sum_{t=1}^{T} N_{t,1} \xrightarrow{P} \frac{1}{2} + (T-1)(1-\frac{\varepsilon}{2}) \qquad \text{and} \qquad \frac{1}{n}\sum_{t=1}^{T} N_{t,0} \xrightarrow{P} \frac{1}{2} + (T-1)\frac{\varepsilon}{2}$$

Thompson Sampling We assume without loss of generality that $\Delta > 0$ and $\pi_1^{(n)} = \frac{1}{2}$. Recall that for Thompson Sampling with independent standard normal priors ($\tilde{\beta}_1, \tilde{\beta}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$) for $a \in [2\colon T]$,

$$\pi_a^{(n)} = \pi_{\min} \vee \left[\pi_{\max} \wedge \mathbb{P}(\tilde{\beta}_1 > \tilde{\beta}_0 \mid H_{a-1}^{(n)})\right]$$

Given the independent standard normal priors on $\tilde{\beta}_1, \tilde{\beta}_0$, we have the following posterior

distribution:

$$\tilde{\beta}_1 - \tilde{\beta}_0 \mid H_{a-1}^{(n)} \sim \mathcal{N}\Bigg(\frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} - \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}}, \\ \frac{\sigma^2(\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}) + \sigma^2(\sigma^2 + \sum_{t=1}^{a-1} N_{t,0})}{(\sigma^2 + \sum_{t=1}^{a-1} N_{t,0})(\sigma^2 + \sum_{t=1}^{a-1} N_{t,1})} \Bigg)$$

$$=: \mathcal{N}\big(\mu_{a-1}^{(n)}, (\sigma_{a-1}^{(n)})^2\big)$$

Thus to show that $\pi_a^{(n)} \xrightarrow{P} \pi_{\max}$ for all $a \in [2\colon T]$, it is sufficient to show that $\mu_{a-1}^{(n)} \xrightarrow{P} \Delta$ and $(\sigma_{a-1}^{(n)})^2 \xrightarrow{P} 0$ for all $a \in [2\colon T]$. By Lemma A.4.1, for all $t \in [1\colon T]$,

$$\frac{N_{t,1}}{\pi_t^{(n)} n} \xrightarrow{P} 1$$

Thus, to show $(\sigma_{a-1}^{(n)})^2 \xrightarrow{P} 0$, it is sufficient to show that

$$\frac{\sigma^2(\sigma^2 + n\sum_{t=1}^{a-1}\pi_t^{(n)}) + \sigma^2(\sigma^2 + n\sum_{t=1}^{a-1}(1-\pi_t^{(n)}))}{(\sigma^2 + n\sum_{t=1}^{a-1}(1-\pi_t^{(n)}))(\sigma^2 + n\sum_{t=1}^{a-1}\pi_t^{(n)})} \xrightarrow{P} 0$$

The above limit holds because $\pi_t^{(n)} \in [\pi_{\min}, \pi_{\max}]$ for $0 < \pi_{\min} \leq \pi_{\max} < 1$ by the clipping condition.

We now show that $\mu_{a-1}^{(n)} \xrightarrow{P} \Delta$, which is equivalent to showing that the following converges in probability to Δ

$$\frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} - \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}}$$

$$= \frac{\sum_{t=1}^{a-1} N_{t,1}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} A_{t,i} R_{t,i}}{\sum_{t=1}^{a-1} N_{t,1}} - \frac{\sum_{t=1}^{a-1} N_{t,0}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}} \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} (1 - A_{t,i}) R_{t,i}}{\sum_{t=1}^{a-1} N_{t,0}}$$

$$= \frac{\sum_{t=1}^{a-1} N_{t,1}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} \left(\beta_1 + \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} A_{t,i} \varepsilon_{t,i}}{\sum_{t=1}^{a-1} N_{t,1}} \right) - \frac{\sum_{t=1}^{a-1} N_{t,0}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}} \left(\beta_0 + \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} (1 - A_{t,i}) \varepsilon_{t,i}}{\sum_{t=1}^{a-1} N_{t,0}} \right) \quad \text{(A.2.9)}$$

Note that

$$\frac{\sum_{t=1}^{a-1} N_{t,1}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} \beta_1 - \frac{\sum_{t=1}^{a-1} N_{t,0}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}} \beta_0 \xrightarrow{P} \Delta \quad \text{(A.2.10)}$$

Equation (A.2.10) above holds by Lemma A.4.1, because

$$\frac{n \sum_{t=1}^{a-1} \pi_t^{(n)}}{\sigma^2 + n \sum_{t=1}^{a-1} \pi_t^{(n)}} \xrightarrow{P} 1 \qquad \frac{n \sum_{t=1}^{a-1} (1 - \pi_t^{(n)})}{\sigma^2 + n \sum_{t=1}^{a-1} (1 - \pi_t^{(n)})} \xrightarrow{P} 1 \quad \text{(A.2.11)}$$

which hold because $\pi_t^{(n)} \in [\pi_{\min}, \pi_{\max}]$ due to our clipping condition.

By Slutsky's Theorem and (A.2.10), to show (A.2.9), it is sufficient to show that

$$\frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} A_{t,i} \varepsilon_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} - \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} (1 - A_{t,i}) \varepsilon_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}} \xrightarrow{P} 0. \quad \text{(A.2.12)}$$

Equation (A.2.12) is equivalent to the following:

$$\sum_{t=1}^{a-1} \frac{\sqrt{N_{t,1}}}{\sigma^2 + \sum_{t'=1}^{a-1} N_{t',1}} \frac{\sum_{i=1}^{n} A_{t,i} \varepsilon_{t,i}}{\sqrt{N_{t,1}}} - \sum_{t=1}^{a-1} \frac{\sqrt{N_{t,0}}}{\sigma^2 + \sum_{t'=1}^{a-1} N_{t',0}} \frac{\sum_{i=1}^{n} (1 - A_{t,i}) \varepsilon_{t,i}}{\sqrt{N_{t,0}}} \xrightarrow{P} 0 \quad \text{(A.2.13)}$$

By Lemma A.4.1, to show (A.2.13) it is sufficient to show that

$$\sum_{t=1}^{a-1} \frac{\sqrt{n\pi_t^{(n)}}}{\sigma^2 + n\sum_{t'=1}^{a-1}\pi_{t'}^{(n)}} \frac{\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{t,1}}} - \sum_{t=1}^{a-1} \frac{\sqrt{n(1-\pi_t^{(n)})}}{\sigma^2 + n\sum_{t'=1}^{a-1}(1-\pi_{t'}^{(n)})} \frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sqrt{N_{t,0}}} \xrightarrow{P} 0 \tag{A.2.14}$$

Since $\pi_t^{(n)} \in [\pi_{\min}, \pi_{\max}]$ due to our clipping condition, the left hand side of (A.2.14) equals the following

$$\sum_{t=1}^{a-1} o_p(1) \frac{\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{t,1}}} - \sum_{t=1}^{a-1} o_p(1) \frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sqrt{N_{t,0}}} \xrightarrow{P} 0$$

The above limit holds by (A.2.8).

Thus, by Slutsky's Theorem and Lemma A.4.1, we have that

$$\frac{1}{n}\sum_{t=1}^{T} N_{t,1} \xrightarrow{P} \frac{1}{2} + (T-1)\pi_{\max} \qquad \text{and} \qquad \frac{1}{n}\sum_{t=1}^{T} N_{t,0} \xrightarrow{P} \frac{1}{2} + (T-1)\pi_{\min} \qquad \square$$

UCB We assume without loss of generality that $\Delta > 0$ and $\pi_1^{(n)} = \frac{1}{2}$. Recall that for UCB, for $a \in [2\colon T]$,

$$\pi_a^{(n)} = \begin{cases} \pi_{\max} & \text{if } U_{a-1,1} > U_{a-1,0} \\ 1 - \pi_{\max} & \text{otherwise} \end{cases}$$

where we define the upper confidence bounds U for any confidence level δ with $0 < \delta < 1$ as follows:

$$U_{a-1,1} = \begin{cases} \infty & \text{if } \sum_{t=1}^{a-1} N_{t,1} = 0 \\ \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sum_{t=1}^{a-1} N_{t,1}} + \sqrt{\frac{2\log 1/\delta}{\sum_{t=1}^{a-1} N_{t,1}}} & \text{otherwise} \end{cases}$$

$$U_{a-1,0} = \begin{cases} \infty & \text{if } N_{1,0} = 0 \\ \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} (1-A_{t,i})R_{t,i}}{\sum_{t=1}^{a-1} N_{t,0}} + \sqrt{\frac{2\log 1/\delta}{\sum_{t=1}^{a-1} N_{t,0}}} & \text{otherwise} \end{cases}$$

Thus to show that $\pi_a^{(n)} \xrightarrow{P} \pi_{\max}$ for all $a \in [2\colon T]$, it is sufficient to show $\mathbb{I}_{(U_{a,1} > U_{a,0})} \xrightarrow{P}$ 1, which is equivalent to showing that the following converges in probability to 1:

$$\mathbb{I}_{(\sum_{t=1}^{a} N_{t,1}>0, \sum_{t=1}^{a} N_{t,0}>0)} \mathbb{I}_{\left(\frac{\sum_{t=1}^{a}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sum_{t=1}^{a} N_{t,1}} + \sqrt{\frac{2\log 1/\delta}{\sum_{t=1}^{a} N_{t,1}}} > \frac{\sum_{t=1}^{a}\sum_{i=1}^{n} (1-A_{t,i})R_{t,i}}{\sum_{t=1}^{a} N_{t,1}} + \sqrt{\frac{2\log 1/\delta}{\sum_{t=1}^{a} N_{t,0}}}\right)}$$
$$+ \mathbb{I}_{(\sum_{t=1}^{a} N_{t,1}=0, \sum_{t=1}^{a} N_{t,0}>0)}$$

$$= \mathbb{I}_{\left((\beta_1-\beta_0) + \frac{\sum_{t=1}^{a}\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sum_{t=1}^{a} N_{t,1}} + \sqrt{\frac{2\log 1/\delta}{\sum_{t=1}^{a} N_{t,1}}} > \frac{\sum_{t=1}^{a}\sum_{i=1}^{n} (1-A_{t,i})\varepsilon_{t,i}}{\sum_{t=1}^{a} N_{t,1}} + \sqrt{\frac{2\log 1/\delta}{\sum_{t=1}^{a} N_{t,0}}}\right)} + o_p(1)$$

Note that to show that the above converges in probability to 1, it is sufficient to show that the following:

$$\frac{\sum_{t=1}^{a}\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sum_{t=1}^{a} N_{t,1}} + \sqrt{\frac{2\log 1/\delta}{\sum_{t=1}^{a} N_{t,0}}} - \frac{\sum_{t=1}^{a}\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sum_{t=1}^{a} N_{t,1}} - \sqrt{\frac{2\log 1/\delta}{\sum_{t=1}^{a} N_{t,1}}} \xrightarrow{P} 0$$

Note for fixed δ, we have that $\frac{2\log 1/\delta}{\sum_{t=1}^{a} N_{t,0}} \xrightarrow{P} 0$, since $\frac{N_{t,0}}{n/2} \xrightarrow{P} 1$. Also note $\frac{\sum_{t=1}^{a}\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sum_{t=1}^{a} N_{t,1}} - \frac{\sum_{t=1}^{a}\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sum_{t=1}^{a} N_{t,1}} \xrightarrow{P} 0$, by the same argument made in the ε-greedy case to show (A.2.5).

Thus, by Slutsky's Theorem and Lemma A.4.1, we have that

$$\frac{1}{n}\sum_{t=1}^{T} N_{t,1} \xrightarrow{P} \frac{1}{2} + (T-1)\pi_{\max} \qquad \text{and} \qquad \frac{1}{n}\sum_{t=1}^{T} N_{t,0} \xrightarrow{P} \frac{1}{2} + (T-1)(1-\pi_{\max})$$

A.3 Non-uniform convergence of the OLS Estimator

Definition A.3.1 (Non-concentration of a sequence of random variables)**.** *For a sequence of random variables* $\{Y_i\}_{i=1}^n$ *on probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, we say* Y_n *does not concentrate if for each* $a \in \mathbb{R}$ *there exists an* $\varepsilon_a > 0$ *with*

$$P\big(\{\omega \in \Omega : \big|Y_n(\omega) - a\big| > \varepsilon_a\}\big) \not\to 0.$$

A.3.1 Thompson Sampling

Proposition A.3.1 (Non-concentration of sampling probabilities under Thompson Sampling)**.** *Under the assumptions of Theorem 3.4.2, the posterior distribution that arm* 1 *is better than arm* 0 *converges as follows:*

$$\mathbb{P}\big(\tilde{\beta}_1 > \tilde{\beta}_0 \,\big|\, H_1^{(n)}\big) \xrightarrow{D} \begin{cases} 1 & \text{if } \Delta > 0 \\ 0 & \text{if } \Delta < 0 \\ \text{Uniform}[0,1] & \text{if } \Delta = 0 \end{cases}$$

Thus, the sampling probabilities $\pi_t^{(n)}$ *do not concentrate when* $\Delta = 0$*.*

Proof: Below, $N_{t,1} = \sum_{i=1}^n A_{t,i}$ and $N_{t,0} = \sum_{i=1}^n (1 - A_{t,i})$. Posterior means:

$$\tilde{\beta}_0 | H_1^{(n)} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n (1 - A_{1,i}) R_{1,i}}{\sigma_a^2 + N_{1,0}}, \frac{\sigma^2}{\sigma_a^2 + N_{0,1}}\right)$$

$$\tilde{\beta}_1 | H_1^{(n)} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{\sigma_a^2 + N_{1,1}}, \frac{\sigma_a^2}{\sigma_a^2 + N_{1,1}}\right)$$

$$\tilde{\beta}_1 - \tilde{\beta}_0 \mid H_1^{(n)} \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

for $\mu_n \triangleq \frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{\sigma_a^2 + N_{1,1}} - \frac{\sum_{i=1}^n (1-A_{1,i}) R_{1,i}}{\sigma_a^2 + N_{1,0}}$ and $\sigma_n^2 \triangleq \frac{\sigma_a^2(\sigma_a^2+N_{1,1}) + \sigma_a^2(\sigma_a^2+N_{1,0})}{(\sigma_a^2+N_{1,0})(\sigma_a^2+N_{1,1})}$.

$$P(\tilde{\beta}_1 > \tilde{\beta}_0 \mid H_1^{(n)}) = P(\tilde{\beta}_1 - \tilde{\beta}_0 > 0 \mid H_1^{(n)}) = P\left(\frac{\tilde{\beta}_1 - \tilde{\beta}_0 - \mu_n}{\sigma_n} > -\frac{\mu_n}{\sigma_n} \,\middle|\, H_1^{(n)} \right)$$

For $Z \sim \mathcal{N}(0,1)$ independent of μ_n, σ_n.

$$= P\left(Z > -\frac{\mu_n}{\sigma_n} \,\middle|\, H_1^{(n)} \right) = P\left(Z < \frac{\mu_n}{\sigma_n} \,\middle|\, H_1^{(n)} \right) = \Phi\left(\frac{\mu_n}{\sigma_n} \,\middle|\, H_1^{(n)} \right)$$

$$\frac{\mu_n}{\sigma_n} = \left(\frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{\sigma_a^2 + N_{1,1}} - \frac{\sum_{i=1}^n (1-A_{1,i}) R_{1,i}}{\sigma_a^2 + N_{1,0}} \right) \sqrt{\frac{(\sigma_a^2 + N_{1,0})(\sigma_a^2 + N_{1,1})}{2\sigma_a^4 + \sigma_a^2 n}}$$

$$= \left(\frac{\beta_1 N_{1,1} + \sum_{i=1}^n A_{1,i} \varepsilon_{1,i}}{\sigma_a^2 + N_{1,1}} - \frac{\beta_0 N_{1,0} + \sum_{i=1}^n (1-A_{1,i}) \varepsilon_{1,i}}{\sigma_a^2 + N_{1,0}} \right) \sqrt{\frac{(\sigma_a^2 + N_{1,0})(\sigma_a^2 + N_{1,1})}{2\sigma_a^4 + \sigma_a^2 n}}$$

$$= \frac{\sum_{i=1}^n A_{1,i} \varepsilon_{1,i}}{\sqrt{N_{1,1}}} \sqrt{\frac{N_{1,1}(\sigma_a^2 + N_{1,0})}{(2\sigma_a^4 + \sigma_a^2 n)(\sigma_a^2 + N_{1,1})}} - \frac{\sum_{i=1}^n (1-A_{1,i}) \varepsilon_{1,i}}{\sqrt{N_{1,0}}} \sqrt{\frac{N_{1,0}(\sigma_a^2 + N_{1,1})}{(2\sigma_a^4 + \sigma_a^2 n)(\sigma_a^2 + N_{1,0})}}$$

$$+ \left(\beta_1 \frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} - \beta_0 \frac{N_{1,0}}{\sigma_a^2 + N_{1,0}} \right) \sqrt{\frac{(\sigma_a^2 + N_{1,0})(\sigma_a^2 + N_{1,1})}{2\sigma_a^4 + \sigma_a^2 n}} =: B_n + C_n$$

Let's first examine C_n. Note that $\beta_1 = \beta_0 + \Delta$, so $\beta_1 \frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} - \beta_0 \frac{N_{1,0}}{\sigma_a^2 + N_{1,0}}$ equals

$$= (\beta_0 + \Delta) \frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} - \beta_0 \frac{N_{1,0}}{\sigma_a^2 + N_{1,0}} = \Delta \frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} + \beta_0 \left(\frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} - \frac{N_{1,0}}{\sigma_a^2 + N_{1,0}} \right)$$

$$= \Delta \frac{N_{1,1}/n}{(\sigma_a^2 + N_{1,1})/n} + \beta_0 \left(\frac{N_{1,1}(\sigma_a^2 + N_{1,0}) - N_{1,0}(\sigma_a^2 + N_{1,1})}{(\sigma_a^2 + N_{1,1})(\sigma_a^2 + N_{1,1})} \right)$$

$$= \Delta \frac{\frac{1}{2} + o(1)}{\frac{1}{2} + o(1)} + \beta_0 \sigma_a^2 \left(\frac{N_{1,1} - N_{1,0}}{(\sigma_a^2 + N_{1,1})(\sigma_a^2 + N_{1,1})} \right) = \Delta[1 + o(1)] + o\left(\frac{1}{n}\right)$$

where the last equality holds by the Strong Law of Large Numbers because

$$\frac{\frac{1}{n^2}(N_{1,1} - N_{1,0})}{\frac{1}{n^2}(\sigma_a^2 + N_{1,1})(\sigma_a^2 + N_{1,1})} = \frac{\frac{1}{n}[\frac{1}{2} - \frac{1}{2} + o(1)]}{[\frac{1}{2} + o(1)][\frac{1}{2} + o(1)]} = \frac{\frac{1}{n} o(1)}{\frac{1}{4} + o(1)} = o\left(\frac{1}{n}\right)$$

Thus,

$$C_n = \left[\Delta[1 + o(1)] + o\left(\frac{1}{n}\right) \right] \sqrt{\frac{(\sigma_a^2 + N_{1,0})(\sigma_a^2 + N_{1,1})}{2\sigma_a^4 + \sigma_a^2 n}}$$

$$= \left[\Delta[1 + o(1)] + o\left(\frac{1}{n}\right) \right] \sqrt{\frac{n[\frac{1}{2} + o(1)][\frac{1}{2} + o(1)]}{o(1) + \sigma_a^2}} = \sqrt{n} \Delta \left[1/(2\sigma_a) + o(1) \right] + o\left(\frac{1}{\sqrt{n}}\right)$$

Let's now examine B_n.

$$\sqrt{\frac{N_{1,1}(\sigma_a^2 + N_{1,0})}{(2\sigma_a^4 + \sigma_a^2 n)(\sigma_a^2 + N_{1,1})}} = \sqrt{\frac{[\frac{1}{2} + o(1)][\frac{1}{2} + o(1)]}{[\sigma_a^2 + o(1)][\frac{1}{2} + o(1)]}} = \sqrt{\frac{1}{2\sigma_a^2}} + o(1)$$

$$\sqrt{\frac{N_{1,0}(\sigma_a^2 + N_{1,1})}{(2\sigma_a^4 + \sigma_a^2 n)(\sigma_a^2 + N_{1,0})}} = \sqrt{\frac{[\frac{1}{2} + o(1)][\frac{1}{2} + o(1)]}{[\sigma_a^2 + o(1)][\frac{1}{2} + o(1)]}} = \sqrt{\frac{1}{2\sigma_a^2}} + o(1)$$

Note that by Theorem 3.5.1, $\left[\frac{1}{\sqrt{N_{1,1}}} \sum_{i=1}^n \varepsilon_{1,i} A_{1,i}, \frac{1}{\sqrt{N_{1,0}}} \sum_{i=1}^n \varepsilon_{1,i}(1 - A_{1,i}) \right] \xrightarrow{D} \mathcal{N}(0, \underline{\mathbf{I}}_2)$. Thus by Slutky's Theorem,

$$\begin{bmatrix} \frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} \sqrt{\frac{N_{1,1}(\sigma_a^2 + N_{1,0})}{(2\sigma_a^4 + \sigma_a^2 n)(\sigma_a^2 + N_{1,1})}} \\ \frac{\sum_{i=1}^n (1 - A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}} \sqrt{\frac{N_{1,0}(\sigma_a^2 + N_{1,1})}{(2\sigma_a^4 + \sigma_a^2 n)(\sigma_a^2 + N_{1,0})}} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} \left[\sqrt{\frac{1}{2\sigma_a^2}} + o(1) \right] \\ \frac{\sum_{i=1}^n (1 - A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}} \left[\sqrt{\frac{1}{2\sigma_a^2}} + o(1) \right] \end{bmatrix} \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{2\sigma_a^2} \underline{\mathbf{I}}_2\right)$$

Thus, we have that, $B_n \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{\sigma_a^2}\right)$. Since we assume that the algorithm's variance is correctly specified, so $\sigma_a^2 = 1$,

$$B_n + C_n \xrightarrow{D} \begin{cases} \infty & \text{if } \Delta > 0 \\ -\infty & \text{if } \Delta < 0 \\ \mathcal{N}(0,1) & \text{if } \Delta = 0 \end{cases}$$

Thus, by continuous mapping theorem,

$$\mathbb{P}\left(\tilde{\beta}_1 > \tilde{\beta}_0 \middle| H_1^{(n)}\right) = \Phi\left(\frac{\mu_n}{\sigma_n}\right) = \Phi(B_n + C_n) \xrightarrow{D} \begin{cases} 1 & \text{if } \Delta > 0 \\ 0 & \text{if } \Delta < 0 \\ \text{Uniform}[0,1] & \text{if } \Delta = 0 \end{cases} \quad \square$$

Proof of Theorem 3.4.2 (Non-uniform convergence of the OLS estimator of the treatment effect for Thompson Sampling): The normalized errors of the OLS estimator for Δ, which are asymptotically normal under i.i.d. sampling are as follows:

$$\sqrt{\frac{(N_{1,1}+N_{2,1})(N_{1,0}+N_{2,0})}{2n}}\left(\hat{\beta}_1^{\text{OLS}} - \hat{\beta}_0^{\text{OLS}} - \Delta\right)$$

$$= \sqrt{\frac{(N_{1,1}+N_{2,1})(N_{1,0}+N_{2,0})}{2n}}\left(\frac{\sum_{t=1}^{2}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{N_{1,1}+N_{2,1}} - \frac{\sum_{t=1}^{2}\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{N_{1,0}+N_{2,0}} - \Delta\right)$$

$$= \sqrt{\frac{(N_{1,1}+N_{2,1})(N_{1,0}+N_{2,0})}{2n}}\left(\beta_1 - \beta_0 - \Delta + \frac{\sum_{t=1}^{2}\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{N_{1,1}+N_{2,1}} - \frac{\sum_{t=1}^{2}\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{N_{1,0}+N_{2,0}}\right)$$

$$= \sqrt{\frac{N_{1,0}+N_{2,0}}{2n}}\frac{\sum_{t=1}^{2}\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{1,1}+N_{2,0}}} - \sqrt{\frac{N_{1,1}+N_{2,1}}{2n}}\frac{\sum_{t=1}^{2}\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{\sqrt{N_{1,0}+N_{2,0}}}$$

$$= [1,-1,1,-1]\begin{bmatrix} \sqrt{\frac{N_{1,0}+N_{2,0}}{2n}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}+N_{2,1}}} \\ \sqrt{\frac{N_{1,1}+N_{2,1}}{2n}}\frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}+N_{2,0}}} \\ \sqrt{\frac{N_{1,0}+N_{2,0}}{2n}}\frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{1,1}+N_{2,1}}} \\ \sqrt{\frac{N_{1,1}+N_{2,1}}{2n}}\frac{\sum_{i=1}^{n}(1-A_{2,i})\varepsilon_{2,i}}{\sqrt{N_{1,0}+N_{2,0}}} \end{bmatrix}$$

$$= [1,-1,1,-1]\begin{bmatrix} \sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}}\sqrt{\frac{N_{1,1}}{n}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{N_{1,1}+N_{2,1}}{2(N_{1,0}+N_{2,0})}}\sqrt{\frac{N_{1,0}}{n}}\frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}} \\ \sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}}\sqrt{\frac{N_{2,1}}{n}}\frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} \\ \sqrt{\frac{N_{1,1}+N_{2,1}}{2(N_{1,0}+N_{2,0})}}\sqrt{\frac{N_{2,0}}{n}}\frac{\sum_{i=1}^{n}(1-A_{2,i})\varepsilon_{2,i}}{\sqrt{N_{2,0}}} \end{bmatrix} \tag{A.3.1}$$

By Theorem 3.5.1, $\left(\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}, \frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}}, \frac{\sum_{i=1}^{n}(1-A_{2,i})\varepsilon_{2,i}}{\sqrt{N_{2,0}}}\right) \xrightarrow{D} \mathcal{N}(0, \underline{\mathbf{I}}_4)$. By Lemma A.4.1 and Slutsky's Theorem, $\sqrt{\frac{2n(N_{1,1}+N_{2,1})}{N_{1,1}(N_{1,0}+N_{2,0})}}\sqrt{\frac{\frac{1}{2}(\frac{1}{2}+[1-\pi_2])}{2(\frac{1}{2}+\pi_2)}} = 1 + o_p(1)$, thus,

$$\sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}}\sqrt{\frac{N_{1,1}}{n}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}$$

$$= \left(\sqrt{\frac{2n(N_{1,1}+N_{2,1})}{N_{1,1}(N_{1,0}+N_{2,0})}}\sqrt{\frac{\frac{1}{2}(\frac{1}{2}+[1-\pi_2])}{2(\frac{1}{2}+\pi_2)}} + o_p(1)\right)\sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}}\sqrt{\frac{N_{1,1}}{n}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}$$

$$= \sqrt{\frac{\frac{1}{2}(\frac{1}{2}+[1-\pi_2])}{2(\frac{1}{2}+\pi_2)}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1)\sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}}\sqrt{\frac{N_{1,1}}{n}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}$$

Note that $\sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}}$ is stochastically bounded because for any $K > 2$,

$$\mathbb{P}\left(\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})} > K\right) \leq \mathbb{P}\left(\frac{n}{N_{1,1}} > K\right) = \mathbb{P}\left(\frac{1}{K} > \frac{N_{1,1}}{n}\right) \to 0$$

where the limit holds by the law of large numbers since $N_{1,1}^{(n)} \sim \text{Binomial}(n, \frac{1}{2})$. Thus, since $\frac{N_{1,1}}{n} \leq 1$ and $\frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} \xrightarrow{D} \mathcal{N}(0,1)$,

$$o_p(1)\sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}}\sqrt{\frac{N_{1,1}}{n}}\frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} = o_p(1)$$

We can perform the above procedure on the other three terms. Thus, equation (A.3.1) is equal to the following:

$$[1,-1,1,-1]\begin{bmatrix} \sqrt{\frac{1/2+1-\pi_2}{4(1/2+\pi_2)}}\frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{1/2+\pi_2}{4(1/2+1-\pi_2)}}\frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}} \\ \sqrt{\frac{(1/2+1-\pi_2)\pi_2}{2(1/2+\pi_2)}}\frac{\sum_{i=1}^{n} A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} \\ \sqrt{\frac{(1/2+\pi_2)(1-\pi_2)}{2(1/2+1-\pi_2)}}\frac{\sum_{i=1}^{n}(1-A_{2,i})\varepsilon_{2,i}}{\sqrt{N_{2,0}}} \end{bmatrix} + o_p(1)$$

Recall that we showed earlier in Proposition A.3.1 that

$$\pi_2^{(n)} = \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(\frac{\mu_n}{\sigma_n}\right)\right] = \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(B_n + C_n\right)\right]$$

$$= \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(\frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{2N_{1,1}}} - \frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{2N_{1,0}}} + \sqrt{n}\Delta\left[\frac{1}{2}+o(1)\right]+o(1)\right)\right]$$

When $\Delta > 0$, $\pi_2^{(n)} \xrightarrow{P} \pi_{\max}$ and when $\Delta < 0$, $\pi_2^{(n)} \xrightarrow{P} \pi_{\min}$. We now consider the

$\Delta = 0$ case.

$$\pi_2^{(n)} = \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(\frac{1}{\sqrt{2}}\left[\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} - \frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}\right] + o(1)\right)\right]$$

$$= \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(\frac{1}{\sqrt{2}}\left[\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} - \frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}\right]\right)\right] + o(1)$$

By Slutsky's Theorem, for $Z_1, Z_2, Z_3, Z_4 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,

$$[1,-1,1,-1]\begin{bmatrix} \sqrt{\frac{1/2+1-\pi_2}{4(1/2+\pi_2)}}\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{1/2+\pi_2}{4(1/2+1-\pi_2)}}\frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}} \\ \sqrt{\frac{(1/2+1-\pi_2)\pi_2}{2(1/2+\pi_2)}}\frac{\sum_{i=1}^n A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} \\ \sqrt{\frac{(1/2+\pi_2)(1-\pi_2)}{2(1/2+1-\pi_2)}}\frac{\sum_{i=1}^n (1-A_{2,i})\varepsilon_{2,i}}{\sqrt{N_{2,0}}} \end{bmatrix} + o_p(1) \overset{D}{\to} [1,-1,1,-1]\begin{bmatrix} \sqrt{\frac{1/2+1-\pi_*}{4(1/2+\pi_*)}}Z_1 \\ \sqrt{\frac{1/2+\pi_*}{4(1/2+1-\pi_*)}}Z_2 \\ \sqrt{\frac{(1/2+1-\pi_*)\pi_*}{2(1/2+\pi_*)}}Z_3 \\ \sqrt{\frac{(1/2+\pi_*)(1-\pi_*)}{2(1/2+1-\pi_*)}}Z_4 \end{bmatrix}$$

$$= \sqrt{\frac{1/2+1-\pi_*}{2(1/2+\pi_*)}}\left(\sqrt{1/2}Z_1 + \sqrt{\pi_*}Z_3\right) - \sqrt{\frac{1/2+\pi_*}{2(1/2+1-\pi_*)}}\left(\sqrt{1/2}Z_2 + \sqrt{1-\pi_*}Z_4\right)$$

$$\text{where } \pi_* = \begin{cases} \pi_{\max} & \text{if } \Delta > 0 \\ \pi_{\min} & \text{if } \Delta < 0 \\ \pi_{\min} \vee (\pi_{\max} \wedge \Phi[\sqrt{1/2}(Z_1 - Z_2)]) & \text{if } \Delta = 0 \end{cases} \quad \square$$

A.3.2 ε-Greedy

Proposition A.3.2 (Non-concentration of the sampling probabilities under zero treatment effect for ε-greedy)**.** *Let $T = 2$ and $\pi_1^{(n)} = \frac{1}{2}$ for all n. We assume that $\{\varepsilon_{t,i}\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and*

$$\pi_2^{(n)} = \begin{cases} 1 - \frac{\varepsilon}{2} & \text{if } \frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) R_{1,i}}{N_{1,0}} \\ \frac{\varepsilon}{2} & \text{otherwise} \end{cases}$$

Thus, the sampling probability $\pi_2^{(n)}$ does not concentrate when $\beta_1 = \beta_0$.

Proof: We define $M_n \triangleq \mathbb{I}_{\left(\frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) R_{1,i}}{N_{1,0}}\right)} = \mathbb{I}_{\left((\beta_1-\beta_0) + \frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{N_{1,0}}\right)}$. Note that when $M_n = 1$, $\pi_2^{(n)} = 1 - \frac{\varepsilon}{2}$ and when $M_n = 0$, $\pi_2^{(n)} = \frac{\varepsilon}{2}$.

When the margin is zero, M_n does not concentrate because for all $N_{1,1}, N_{1,0}$, since $\varepsilon_{1,i} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,

$$\mathbb{P}\left(\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{N_{1,0}}\right) = \mathbb{P}\left(\frac{1}{\sqrt{N_{1,1}}} Z_1 - \frac{1}{\sqrt{N_{1,0}}} Z_2 > 0\right) = \frac{1}{2}$$

for $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Thus, we have shown that $\pi_2^{(n)}$ does not concentrate when $\beta_1 - \beta_0 = 0$. □

Theorem A.3.1 (Non-uniform convergence of the OLS estimator of the treatment effect for ε-greedy)**.** *Assuming the setup and conditions of Proposition A.3.2, and that $\beta_1 = b$, we show that the normalized errors of the OLS estimator converges in distribution as follows:*

$$\sqrt{N_{1,1} + N_{2,1}}\left(\hat{\beta}_1^{\text{OLS}} - b\right) \xrightarrow{D} Y$$

$$Y = \begin{cases} Z_1 & \text{if } \beta_1 - \beta_0 \neq 0 \\ \sqrt{\frac{1}{3-\varepsilon}}\big(Z_1 - \sqrt{2-\varepsilon}Z_3\big)\mathbb{I}_{(Z_1>Z_2)} + \sqrt{\frac{1}{1+\varepsilon}}\big(Z_1 - \sqrt{\varepsilon}Z_3\big)\mathbb{I}_{(Z_1<Z_2)} & \text{if } \beta_1 - \beta_0 = 0 \end{cases}$$

for $Z_1, Z_2, Z_3 \overset{i.i.d.}{\sim} N(0,1)$. *Note the* $\beta_1 - \beta_0 = 0$ *case,* Y *is non-normal.*

Proof: The normalized errors of the OLS estimator for β_1 are

$$\sqrt{N_{1,1}+N_{2,1}}\left(\frac{\sum_{t=1}^{2}\sum_{i=1}^{n}A_{t,i}R_{t,i}}{N_{1,1}+N_{2,1}} - b\right) = \frac{\sum_{t=1}^{2}\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{1,1}+N_{2,1}}}$$

$$= [1,1]\begin{bmatrix}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}+N_{2,1}}} \\ \frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{1,1}+N_{2,1}}}\end{bmatrix} = [1,1]\begin{bmatrix}\sqrt{\frac{N_{1,1}}{N_{1,1}+N_{2,1}}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{N_{2,1}}{N_{1,1}+N_{2,1}}}\frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}}\end{bmatrix}$$

By Slutsky's Theorem and Lemma A.4.1, $\left(\sqrt{\frac{1/2}{1/2+\pi_2^{(n)}}}\sqrt{\frac{N_{1,1}+N_{2,1}}{N_{1,1}}}, \sqrt{\frac{\pi_2^{(n)}}{1/2+\pi_2^{(n)}}}\sqrt{\frac{N_{1,1}+N_{2,1}}{N_{2,1}}}\right) \xrightarrow{P} (1,1)$, so

$$= [1,1]\begin{bmatrix}\left(\sqrt{\frac{1/2}{1/2+\pi_2^{(n)}}}\sqrt{\frac{N_{1,1}+N_{2,1}}{N_{1,1}}} + o_p(1)\right)\sqrt{\frac{N_{1,1}}{N_{1,1}+N_{2,1}}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \left(\sqrt{\frac{\pi_2^{(n)}}{1/2+\pi_2^{(n)}}}\sqrt{\frac{N_{1,1}+N_{2,1}}{N_{2,1}}} + o_p(1)\right)\sqrt{\frac{N_{2,1}}{N_{1,1}+N_{2,1}}}\frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}}\end{bmatrix}$$

$$= [1,1]\begin{bmatrix}\sqrt{\frac{1/2}{1/2+\pi_2^{(n)}}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\pi_2^{(n)}}{1/2+\pi_2^{(n)}}}\frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1)\end{bmatrix}$$

The last equality holds because by Theorem 3.5.1, $\left(\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}}\right) \xrightarrow{D} \mathcal{N}(0, \underline{\mathbf{I}}_2)$.

Let's define $M_n \triangleq \mathbb{I}_{\left(\frac{\sum_{i=1}^{n}A_{1,i}R_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^{n}(1-A_{1,i})R_{1,i}}{N_{1,0}}\right)} = \mathbb{I}_{\left((\beta_1-\beta_0)+\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{N_{1,0}}\right)}$.

Note that when $M_n = 1$, $\pi_2^{(n)} = 1 - \frac{\varepsilon}{2}$ and when $M_n = 0$, $\pi_2^{(n)} = \frac{\varepsilon}{2}$.

$$M_n = \mathbb{I}_{\left((\beta_1-\beta_0)+\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{N_{1,0}}\right)} = \mathbb{I}_{\left(\sqrt{N_{1,0}}(\beta_1-\beta_0)+\sqrt{\frac{N_{1,0}}{N_{1,1}}}\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} > \frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}\right)}$$

$$= \mathbb{I}_{\left(\sqrt{N_{1,0}}(\beta_1-\beta_0)+[1+o_p(1)]\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} > \frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}\right)}$$

where the last equality holds because $\sqrt{\frac{N_{1,0}}{N_{1,1}}} \xrightarrow{P} 1$ by Lemma A.4.1, Slutsky's Theorem, and continuous mapping theorem. Thus, by Proposition A.3.2,

$$M^{(n)} \xrightarrow{P} \begin{cases} 1 & \text{if } \beta_1 - \beta_0 > 0 \\ 0 & \text{if } \beta_1 - \beta_0 < 0 \\ \text{does not concentrate} & \text{if } \beta_1 - \beta_0 = 0 \end{cases}$$

Note that

$$\begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\pi_2^{(n)}}}\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\pi_2^{(n)}}{\frac{1}{2}+\pi_2^{(n)}}}\frac{\sum_{i=1}^n A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+1-\frac{\varepsilon}{2}}}\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{1-\varepsilon/2}{\frac{1}{2}+1-\frac{\varepsilon}{2}}}\frac{\sum_{i=1}^n A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} M_n + \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\frac{\varepsilon}{2}}}\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\frac{\varepsilon}{2}}{\frac{1}{2}+\frac{\varepsilon}{2}}}\frac{\sum_{i=1}^n A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} (1 - M_n)$$

Also note by Theorem 3.5.1, $\left(\frac{\sum_{i=1}^n A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^n (1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}, \frac{\sum_{i=1}^n A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}}, \frac{\sum_{i=1}^n (1-A_{2,i})\varepsilon_{2,i}}{\sqrt{N_{2,1}}}\right) \xrightarrow{D} \mathcal{N}(0, \underline{\mathbf{I}}_4)$.

When $\beta_1 > \beta_0$, $M_n \xrightarrow{P} 1$ and when $\beta_1 < \beta_0$, $M_n \xrightarrow{P} 0$; in both these cases the normalized errors are asymptotically normal. We now focus on the case that $\beta_1 = \beta_0$. By continuous

mapping theorem and Slutsky's theorem for $Z_1, Z_2, Z_3, Z_4 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,

$$= [1,1] \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+1-\frac{\varepsilon}{2}}} \frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{1-\varepsilon/2}{\frac{1}{2}+1-\frac{\varepsilon}{2}}} \frac{\sum_{i=1}^{n} A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} \mathbb{I}_{\left([1+o(1)]\frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} > \frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}\right)}$$

$$+[1,1] \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\frac{\varepsilon}{2}}} \frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\frac{\varepsilon}{2}}{\frac{1}{2}+\frac{\varepsilon}{2}}} \frac{\sum_{i=1}^{n} A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} \left(1 - \mathbb{I}_{\left([1+o(1)]\frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} > \frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}\right)}\right)$$

$$\overset{D}{\to} [1,1] \begin{bmatrix} \sqrt{\frac{1/2}{1/2+1-\varepsilon/2}} Z_1 \\ \sqrt{\frac{1-\varepsilon/2}{1/2+1-\varepsilon/2}} Z_3 \end{bmatrix} \mathbb{I}_{(Z_1>Z_2)} + [1,1] \begin{bmatrix} \sqrt{\frac{1/2}{1/2+\varepsilon/2}} Z_1 \\ \sqrt{\frac{\varepsilon/2}{1/2+\varepsilon/2}} Z_3 \end{bmatrix} \mathbb{I}_{(Z_1<Z_2)}$$

$$= \left(\sqrt{\frac{1}{3-\varepsilon}} Z_1 + \sqrt{\frac{2-\varepsilon}{3-\varepsilon}} Z_3\right) \mathbb{I}_{(Z_1>Z_2)} + \left(\sqrt{\frac{1}{1+\varepsilon}} Z_1 + \sqrt{\frac{\varepsilon}{1+\varepsilon}} Z_3\right) \mathbb{I}_{(Z_1<Z_2)}$$

Thus,

$$\frac{\sum_{t=1}^{2}\sum_{i=1}^{n} A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{1,1}+N_{2,1}}} \overset{D}{\to} Y$$

$$Y \triangleq \begin{cases} \sqrt{\frac{1}{3-\varepsilon}}\left(Z_1 - \sqrt{2-\varepsilon} Z_3\right) & \text{if } \beta_1 - \beta_0 > 0 \\ \sqrt{\frac{1}{1+\varepsilon}}\left(Z_1 - \sqrt{\varepsilon} Z_3\right) & \text{if } \beta_1 - \beta_0 < 0 \\ \sqrt{\frac{1}{3-\varepsilon}}\left(Z_1 - \sqrt{2-\varepsilon} Z_3\right)\mathbb{I}_{(Z_1>Z_2)} + \sqrt{\frac{1}{1+\varepsilon}}\left(Z_1 - \sqrt{\varepsilon} Z_3\right)\mathbb{I}_{(Z_1<Z_2)} & \text{if } \beta_1 - \beta_0 = 0 \end{cases} \square$$

A.3.3 UCB

Theorem A.3.2 (Asymptotic non-Normality under zero treatment effect for clipped UCB). *Let $T = 2$ and $\pi_1^{(n)} = \frac{1}{2}$ for all n. We assume that $\{\varepsilon_{t,i}\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and*

$$\pi_2^{(n)} = \begin{cases} \pi_{\max} & \text{if } U_1 > U_0 \\ 1 - \pi_{\max} & \text{otherwise} \end{cases}$$

where we define the upper confidence bounds U for any confidence level δ with $0 < \delta < 1$ as follows:

$$U_1 = \begin{cases} \infty & \text{if } N_{1,1} = 0 \\ \frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{N_{1,1}} + \sqrt{\frac{2 \log 1/\delta}{N_{1,1}}} & \text{otherwise} \end{cases}$$

$$U_0 = \begin{cases} \infty & \text{if } N_{1,0} = 0 \\ \frac{\sum_{i=1}^n (1-A_{1,i}) R_{1,i}}{N_{1,1}} + \sqrt{\frac{2 \log 1/\delta}{N_{1,0}}} & \text{otherwise} \end{cases}$$

Assuming above conditions, and that $\beta_1 = b$, we show that the normalized errors of the OLS estimator converges in distribution as follows:

$$\sqrt{N_{1,1} + N_{2,1}}\big(\hat{\beta}_1^{\text{OLS}} - b\big) \xrightarrow{D} Y$$

$$Y = \begin{cases} Z_1 & \text{if } \Delta = 0 \\ \big(\sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\pi_{\max}}} Z_1 + \sqrt{\frac{\pi_{\max}}{\frac{1}{2}+\pi_{\max}}} Z_3\big) \mathbb{I}_{(Z_1 > Z_2)} + \big(\sqrt{\frac{\frac{1}{2}}{\frac{3}{2}-\pi_{\max}}} Z_1 + \sqrt{\frac{1-\pi_{\max}}{\frac{3}{2}-\pi_{\max}}} Z_3\big) \mathbb{I}_{(Z_1 < Z_2)} & \text{if } \Delta = 0 \end{cases}$$

for $Z_1, Z_2, Z_3 \overset{i.i.d.}{\sim} N(0,1)$. Note the $\Delta \triangleq \beta_1 - \beta_0 = 0$ case, Y is non-normal.

PROOF: The proof is very similar to that of asymptotic non-normality result for ε-Greedy. By the same arguments made as in the ε-Greedy case, we have that

$$\sqrt{N_{1,1}+N_{2,1}}\left(\frac{\sum_{t=1}^{2}\sum_{i=1}^{n}A_{t,i}R_{t,i}}{N_{1,1}+N_{2,1}}-b\right)=[1,1]\begin{bmatrix}\sqrt{\frac{1/2}{1/2+\pi_2^{(n)}}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}+o_p(1)\\ \sqrt{\frac{\pi_2^{(n)}}{1/2+\pi_2^{(n)}}}\frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}}+o_p(1)\end{bmatrix}$$

Assuming $n \geq 1$, we then define

$$M_n \triangleq \mathbb{I}_{(U_1>U_0)}$$

$$=\mathbb{I}_{(N_{1,1}>0,N_{1,0}>0)}\mathbb{I}\left(\frac{\sum_{i=1}^{n}A_{1,i}R_{1,i}}{N_{1,1}}+\sqrt{\frac{2\log 1/\delta}{N_{1,1}}}>\frac{\sum_{i=1}^{n}(1-A_{1,i})R_{1,i}}{N_{1,1}}+\sqrt{\frac{2\log 1/\delta}{N_{1,0}}}\right)+\mathbb{I}_{(N_{1,1}=0,N_{1,0}>0)}$$

$$=\mathbb{I}_{(N_{1,1}>0,N_{1,0}>0)}\mathbb{I}\left((\beta_1-\beta_0)+\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{N_{1,1}}+\sqrt{\frac{2\log 1/\delta}{N_{1,1}}}>\frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{N_{1,1}}+\sqrt{\frac{2\log 1/\delta}{N_{1,0}}}\right)+\mathbb{I}_{(N_{1,1}=0,N_{1,0}>0)}$$

$$=\mathbb{I}_{(N_{1,1}>0,N_{1,0}>0)}\mathbb{I}\left(\sqrt{N_{1,0}}(\beta_1-\beta_0)+\sqrt{\frac{N_{1,0}}{N_{1,1}}}\left[\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}+\sqrt{2\log 1/\delta}\right]>\frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,1}}}+\sqrt{2\log 1/\delta}\right)$$
$$+\mathbb{I}_{(N_{1,1}=0,N_{1,0}>0)}$$

Note that $\frac{N_{1,0}}{N_{1,1}} \xrightarrow{P} 1$ by Lemma A.4.1. Thus by Slutsky's Theorem and continuous mapping theorem,

$$=\mathbb{I}\left(\sqrt{N_{1,0}}(\beta_1-\beta_0)+[1+o_p(1)]\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}+o_p(1)>\frac{\sum_{i=1}^{n}(1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,1}}}\right)+o_p(1) \tag{A.3.2}$$

Note that

$$\begin{bmatrix}\sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\pi_2^{(n)}}}\frac{\sum_{i=1}^{n}A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}+o_p(1)\\ \sqrt{\frac{\pi_2^{(n)}}{\frac{1}{2}+\pi_2^{(n)}}}\frac{\sum_{i=1}^{n}A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}}+o_p(1)\end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\pi_{\max}}} \frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\pi_{\max}}{\frac{1}{2}+\pi_{\max}}} \frac{\sum_{i=1}^{n} A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} M_n + \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+1-\pi_{\max}}} \frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{1-\pi_{\max}}{\frac{1}{2}+1-\pi_{\max}}} \frac{\sum_{i=1}^{n} A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} (1 - M_n)$$

Let $(Z_1^{(n)}, Z_2^{(n)}, Z_3^{(n)}, Z_4^{(n)}) \triangleq \left(\frac{\sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^{n} (1-A_{1,i})\varepsilon_{1,i}}{\sqrt{N_{1,0}}}, \frac{\sum_{i=1}^{n} A_{2,i}\varepsilon_{2,i}}{\sqrt{N_{2,1}}}, \frac{\sum_{i=1}^{n} (1-A_{2,i})\varepsilon_{2,i}}{\sqrt{N_{2,1}}} \right)$. Note that by Theorem 3.5.1, $(Z_1^{(n)}, Z_2^{(n)}, Z_3^{(n)}, Z_4^{(n)}) \xrightarrow{D} \mathcal{N}(0, \underline{\mathbf{I}}_4)$.

When $\beta_1 > \beta_0$, $M_n \xrightarrow{P} 1$ and when $\beta_1 < \beta_0$, $M_n \xrightarrow{P} 0$; in both these cases the normalized errors are asymptotically normal. We now focus on the case that $\beta_1 = \beta_0$. By continuous mapping theorem and Slutsky's theorem,

$$\begin{aligned} &= [1,1] \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\pi_{\max}}} Z_1^{(n)} + o_p(1) \\ \sqrt{\frac{\pi_{\max}}{\frac{1}{2}+\pi_{\max}}} Z_3^{(n)} + o_p(1) \end{bmatrix} \left[\mathbb{I}_{\left([1+o_p(1)] Z_1^{(n)} + o_p(1) > Z_2^{(n)} \right)} + o_p(1) \right] \\ &\qquad + [1,1] \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+1-\pi_{\max}}} Z_1^{(n)} + o_p(1) \\ \sqrt{\frac{1-\pi_{\max}}{\frac{1}{2}+1-\pi_{\max}}} Z_3^{(n)} + o_p(1) \end{bmatrix} \left[1 - \mathbb{I}_{\left([1+o_p(1)] Z_1^{(n)} + o_p(1) > Z_2^{(n)} \right)} + o_p(1) \right] \Bigg) \\ &\xrightarrow{D} [1,1] \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\pi_{\max}}} Z_1 \\ \sqrt{\frac{\pi_{\max}}{\frac{1}{2}+\pi_{\max}}} Z_3 \end{bmatrix} \mathbb{I}_{(Z_1 > Z_2)} + [1,1] \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+1-\pi_{\max}}} Z_1 \\ \sqrt{\frac{1-\pi_{\max}}{\frac{1}{2}+1-\pi_{\max}}} Z_3 \end{bmatrix} \mathbb{I}_{(Z_1 < Z_2)} \end{aligned}$$

$$= \left(\sqrt{\frac{\frac{1}{2}}{\frac{1}{2}+\pi_{\max}}} Z_1 + \sqrt{\frac{\pi_{\max}}{\frac{1}{2}+\pi_{\max}}} Z_3 \right) \mathbb{I}_{(Z_1 > Z_2)} + \left(\sqrt{\frac{\frac{1}{2}}{\frac{3}{2}-\pi_{\max}}} Z_1 + \sqrt{\frac{1-\pi_{\max}}{\frac{3}{2}-\pi_{\max}}} Z_3 \right) \mathbb{I}_{(Z_1 < Z_2)}. \quad \square$$

Note that (A.3.2) implies that if $\beta_1 = \beta_0$, that $\pi_2^{(n)}$ will not concentrate.

A.4 Batched OLS Estimator Asymptotic Normality: Multi-Arm Bandits

Theorem 3.5.1 (Asymptotic normality of Batched OLS estimator for multi-arm bandits) *Assuming Conditions A.2.1 (weak moments) and 3.4.3 (conditionally i.i.d. actions), and a clipping rate of $f(n) = \omega(\frac{1}{n})$ (Definition 3.3.1),*

$$\begin{bmatrix} \begin{bmatrix} N_{1,0} & 0 \\ 0 & N_{1,1} \end{bmatrix}^{1/2} (\hat{\beta}_1^{\text{BOLS}} - \beta_1) \\ \begin{bmatrix} N_{2,0} & 0 \\ 0 & N_{2,1} \end{bmatrix}^{1/2} (\hat{\beta}_2^{\text{BOLS}} - \beta_2) \\ \vdots \\ \begin{bmatrix} N_{T,0} & 0 \\ 0 & N_{T,1} \end{bmatrix}^{1/2} (\hat{\beta}_T^{\text{BOLS}} - \beta_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{2T})$$

where $\beta_t = (\beta_{t,0}, \beta_{t,1})$, $N_{t,1} = \sum_{i=1}^n A_{t,i}$, and $N_{t,0} = \sum_{i=1}^n (1 - A_{t,i})$. Note in the body of this paper, we state Theorem 3.5.1 with conditions that are are sufficient for the weaker conditions we use here.

Lemma A.4.1. *Assuming the conditions of Theorem 3.5.1, for any batch $t \in [1\colon T]$,*

$$\frac{N_{t,1}}{n\pi_t^{(n)}} = \frac{\sum_{i=1}^n A_{t,i}}{n\pi_t^{(n)}} \xrightarrow{P} 1 \quad \text{and} \quad \frac{N_{t,0}}{n(1-\pi_t^{(n)})} = \frac{\sum_{i=1}^n (1 - A_{t,i})}{n(1-\pi_t^{(n)})} \xrightarrow{P} 1$$

PROOF OF LEMMA A.4.1: Note that showing $\frac{N_{t,1}}{n\pi_t^{(n)}} \xrightarrow{P} 1$ is equivalent to showing that $\frac{1}{n\pi_t^{(n)}} \sum_{i=1}^{n} (A_{t,i} - \pi_t^{(n)}) \xrightarrow{P} 0$. Let $\varepsilon > 0$.

$$\mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^{n} (A_{t,i} - \pi_t^{(n)}) \right| > \varepsilon \right)$$

$$= \mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^{n} (A_{t,i} - \pi_t^{(n)}) \right| \left[\mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} + \mathbb{I}_{(\pi_t^{(n)} \notin [f(n), 1-f(n)])} \right] > \varepsilon \right)$$

$$\leq \mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^{n} (A_{t,i} - \pi_t^{(n)}) \right| \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} > \frac{\varepsilon}{2} \right)$$
$$+ \mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^{n} (A_{t,i} - \pi_t^{(n)}) \right| \mathbb{I}_{(\pi_t^{(n)} \notin [f(n), 1-f(n)])} > \frac{\varepsilon}{2} \right)$$

Since by our clipping assumption, $\mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \xrightarrow{P} 1$, the second probability in the summation above converges to 0 as $n \to \infty$. We will now show that the first probability in the summation above also goes to zero. Note that $\mathbb{E}\left[\frac{1}{n\pi_t^{(n)}} \sum_{i=1}^{n} (A_{t,i} - \pi_t^{(n)}) \right] = \mathbb{E}\left[\frac{1}{n\pi_t^{(n)}} \sum_{i=1}^{n} (\mathbb{E}[A_{t,i} | H_{t-1}^{(n)}] - \pi_t^{(n)}) \right] = 0$. So by Chebychev inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^{n} (A_{t,i} - \pi_t^{(n)}) \right| \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} > \varepsilon \right)$$

$$\leq \frac{1}{\varepsilon^2 n^2} \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} \left(\sum_{i=1}^{n} (A_{t,i} - \pi_t^{(n)}) \right)^2 \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \right]$$

$$\leq \frac{1}{\varepsilon^2 n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} (A_{t,i} - \pi_t^{(n)})(A_{t,j} - \pi_t^{(n)}) \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \right]$$

$$= \frac{1}{\varepsilon^2 n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \mathbb{E}\left[A_{t,i} A_{t,j} - \pi_t^{(n)} (A_{t,i} + A_{t,j}) + (\pi_t^{(n)})^2 \middle| H_{t-1}^{(n)} \right] \right]$$

$$= \frac{1}{\varepsilon^2 n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \left(\mathbb{E}\left[A_{t,i} A_{t,j} \middle| H_{t-1}^{(n)} \right] - (\pi_t^{(n)})^2 \right) \right] \tag{A.4.1}$$

Note if $i \neq j$, since $A_{t,i} \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi_t^{(n)})$, $\mathbb{E}[A_{t,i} A_{t,j} | H_{t-1}^{(n)}] = \mathbb{E}[A_{t,i} | H_{t-1}^{(n)}] E[A_{t,j} | H_{t-1}^{(n)}] = (\pi_t^{(n)})^2$, so (A.4.1) above equals the following

$$= \frac{1}{\varepsilon^2 n^2} \sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \left(\mathbb{E}\left[A_{t,i} \middle| H_{t-1}^{(n)} \right] - (\pi_t^{(n)})^2 \right) \right]$$

$$= \frac{1}{\varepsilon^2 n^2} \sum_{i=1}^{n} \mathbb{E}\left[\frac{1 - \pi_t^{(n)}}{\pi_t^{(n)}} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \right] = \frac{1}{\varepsilon^2 n} \mathbb{E}\left[\frac{1 - \pi_t^{(n)}}{\pi_t^{(n)}} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \right]$$

$$\leq \frac{1}{\varepsilon^2 n} \frac{1}{f(n)} \to 0$$

where the limit holds because we assume $f(n) = \omega(\frac{1}{n})$ so $f(n) n \to \infty$. We can make a very similar argument for $\frac{N_{t,0}}{n(1-\pi_t^{(n)})} \xrightarrow{P} 1$. □

Proof for Theorem 3.5.1 (Asymptotic normality of Batched OLS estimator for multi-arm bandits): For readability, for this proof we drop the (n) superscript on $\pi_t^{(n)}$. Note that

$$\begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{1/2} (\hat{\beta}_t^{\text{BOLS}} - \beta_t) = \begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} 1 - A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i}.$$

We want to show that

$$\begin{bmatrix} \begin{bmatrix} N_{0,1} & 0 \\ 0 & N_{1,1} \end{bmatrix}^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \varepsilon_{1,i} \\ \begin{bmatrix} N_{0,2} & 0 \\ 0 & N_{1,2} \end{bmatrix}^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \varepsilon_{2,i} \\ \vdots \\ \begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \varepsilon_{T,i} \end{bmatrix} = \begin{bmatrix} N_{0,1}^{-1/2} \sum_{i=1}^{n} (1-A_{1,i})\varepsilon_{1,i} \\ N_{1,1}^{-1/2} \sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i} \\ N_{0,2}^{-1/2} \sum_{i=1}^{n} (1-A_{2,i})\varepsilon_{2,i} \\ N_{1,2}^{-1/2} \sum_{i=1}^{n} A_{2,i}\varepsilon_{2,i} \\ \vdots \\ N_{t,0}^{-1/2} \sum_{i=1}^{n} (1-A_{T,i})\varepsilon_{T,i} \\ N_{t,1}^{-1/2} \sum_{i=1}^{n} A_{T,i}\varepsilon_{T,i} \end{bmatrix}$$

$$\xrightarrow{D} \mathcal{N}(0, \sigma^2 \mathbf{I}_{2T}).$$

By Lemma A.4.1 and Slutsky's Theorem it is sufficient to show that as $n \to \infty$,

$$\begin{bmatrix} \frac{1}{\sqrt{n(1-\pi_1)}} \sum_{i=1}^{n} (1-A_{1,i})\varepsilon_{1,i} \\ \frac{1}{\sqrt{n\pi_1}} \sum_{i=1}^{n} A_{1,i}\varepsilon_{1,i} \\ \frac{1}{\sqrt{n(1-\pi_2)}} \sum_{i=1}^{n} (1-A_{2,i})\varepsilon_{2,i} \\ \frac{1}{\sqrt{n\pi_2}} \sum_{i=1}^{n} A_{2,i}\varepsilon_{2,i} \\ \vdots \\ \frac{1}{\sqrt{n(1-\pi_T)}} \sum_{i=1}^{n} (1-A_{T,i})\varepsilon_{T,i} \\ \frac{1}{\sqrt{n\pi_T}} \sum_{i=1}^{n} A_{T,i}\varepsilon_{T,i} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}} \begin{bmatrix} 1-\pi_{1,1} & 0 \\ 0 & \pi_{1,1} \end{bmatrix}^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \varepsilon_{1,i} \\ \frac{1}{\sqrt{n}} \begin{bmatrix} 1-\pi_2^{(n)} & 0 \\ 0 & \pi_2^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \varepsilon_{2,i} \\ \vdots \\ \frac{1}{\sqrt{n}} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \varepsilon_{T,i} \end{bmatrix}$$

$$\xrightarrow{D} \mathcal{N}(0, \sigma^2 \mathbf{I}_{2T})$$

By Cramer-Wold device, it is sufficient to show that for any fixed vector $\mathbf{c} \in \mathbb{R}^{2T}$ s.t. $\|\mathbf{c}\|_2 = 1$ that as $n \to \infty$,

$$\mathbf{c}^\top \begin{bmatrix} n^{-1/2} \begin{bmatrix} 1-\pi_{1,1} & 0 \\ 0 & \pi_{1,1} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \varepsilon_{1,i} \\ n^{-1/2} \begin{bmatrix} 1-\pi_2^{(n)} & 0 \\ 0 & \pi_2^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \varepsilon_{2,i} \\ \vdots \\ n^{-1/2} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \varepsilon_{T,i} \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Let us break up $\mathbf{c}$ so that $\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T]^\top \in \mathbb{R}^{2T}$ with $\mathbf{c}_t \in \mathbb{R}^2$ for $t \in [1\colon T]$. The above is equivalent to

$$\sum_{t=1}^T n^{-1/2} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Let us define $Y_{t,i} := n^{-1/2} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_{t,i} & 0 \\ 0 & \pi_{t,i} \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i}$.

The sequence $\{Y_{1,1}, Y_{1,2}, ..., Y_{1,n}, ..., Y_{T,1}, Y_{T,2}, ..., Y_{T,n}\}$ is a martingale with respect to sequence of histories $\{H_t^{(n)}\}_{t=1}^T$, since

$$\mathbb{E}[Y_{t,i} | H_{t-1}^{(n)}] = n^{-1/2} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \mathbb{E}\left[\begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i} \middle| H_{t-1}^{(n)} \right]$$

$$= n^{-1/2}\mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \mathbb{E}\left[\begin{bmatrix} (1-\pi_t^{(n)})E[\varepsilon_{t,i}|H_{t-1}^{(n)}, A_{t,i}=0] \\ \pi_{t,i}E[\varepsilon_{t,i}|H_{t-1}^{(n)}, A_{t,i}=1] \end{bmatrix} \middle| H_{t-1}^{(n)} \right] = 0$$

for all $i \in [1\colon n]$ and all $t \in [1\colon T]$. We then apply[29] martingale central limit theorem to $Y_{t,i}$ to show the desired result (see the proof of Theorem 3.4.1 in Appendix A.2 for the statement of the martingale CLT conditions).

Condition(a): Martingale Condition The first condition holds because $\mathbb{E}[Y_{t,i}|H_{t-1}^{(n)}] = 0$ for all $i \in [1\colon n]$ and all $t \in [1\colon T]$.

Condition(b): Conditional Variance

$$\sum_{t=1}^{T}\sum_{i=1}^{n} E[Y_{n,t,i}^2|H_{t-1}^{(n)}] = \sum_{t=1}^{T}\sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i} \right)^2 \middle| H_{t-1}^{(n)} \right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{n}\mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} & 0 \\ 0 & A_{t,i} \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \mathbf{c}_t\varepsilon_{t,i}^2 \middle| H_{t-1}^{(n)} \right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{1}{n}\mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \begin{bmatrix} \mathbb{E}[(1-A_{t,i})\varepsilon_{t,i}^2|H_{t-1}^{(n)}] & 0 \\ 0 & \mathbb{E}[A_{t,i}\varepsilon_{t,i}^2|H_{t-1}^{(n)}] \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \mathbf{c}_t$$

Since $\mathbb{E}[A_{t,i}\varepsilon_{t,i}^2|H_{t-1}^{(n)}] = \pi_t^{(n)}\mathbb{E}[\varepsilon_{t,i}^2|H_{t-1}^{(n)}, A_{t,i}=1] = \sigma^2\pi_t$ and $\mathbb{E}[(1-A_{t,i})\varepsilon_{t,i}^2|H_{t-1}^{(n)}] =$

$$(1-\pi_t)\mathbb{E}[\varepsilon_{t,i}^2|H_{t-1}^{(n)}, A_{t,i}=0] = \sigma^2(1-\pi_t),$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n} n^{-1}\mathbf{c}_t^\top \mathbf{c}_t\sigma^2 = \sum_{t=1}^{T}\mathbf{c}_t^\top \mathbf{c}_t\sigma^2 = \sigma^2$$

CONDITION(C): LINDEBERG CONDITION Let $\delta > 0$.

$$\sum_{t=1}^{T}\sum_{i=1}^{n} E\left[Y_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2>\delta^2)}\middle|H_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\left[\left(n^{-1/2}\mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i}\right)^2 \mathbb{I}_{(Y_{t,i}^2>\delta^2)}\middle|H_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} & 0 \\ 0 & A_{t,i} \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \mathbf{c}_t\varepsilon_{t,i}^2\mathbb{I}_{(Y_{t,i}^2>\delta^2)}\middle|H_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} \mathbb{E}[(1-A_{t,i})\varepsilon_{t,i}^2\mathbb{I}_{(Y_{t,i}^2>\delta^2)}|H_{t-1}^{(n)}] & 0 \\ 0 & \mathbb{E}[A_{t,i}\varepsilon_{t,i}^2\mathbb{I}_{(Y_{t,i}^2>\delta^2)}|H_{t-1}^{(n)}] \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-\frac{1}{2}} \mathbf{c}_t$$

Note for $\mathbf{c}_t = [c_{t,0}, c_{t,1}]^\top$, $\mathbb{E}\left[(1-A_{t,i})\varepsilon_{t,i}^2\mathbb{I}_{(Y_{t,i}^2>\delta^2)}\middle|H_{t-1}^{(n)}\right] = \mathbb{E}\left[\varepsilon_{t,i}^2\mathbb{I}_{\left(\frac{c_{t,0}^2}{1-\pi_t^{(n)}}\varepsilon_{t,i}^2>n\delta^2\right)}\middle|H_{t-1}^{(n)}, A_{t,i}=0\right](1-\pi_t)$ and $\mathbb{E}\left[A_{t,i}\varepsilon_{t,i}^2\mathbb{I}_{(Y_{t,i}^2>\delta^2)}\middle|H_{t-1}^{(n)}\right] = \mathbb{E}\left[\varepsilon_{t,i}^2\mathbb{I}_{\left(\frac{c_{t,1}^2}{\pi_t^{(n)}}\varepsilon_{t,i}^2>n\delta^2\right)}\middle|H_{t-1}^{(n)}, A_{t,i}=1\right]\pi_t$. Thus, we

have that

$$= \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} c_{t,0}^2 \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{\left(\varepsilon_{t,i}^2 > \frac{n\delta^2(1-\pi_t)}{c_{t,0}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 0 \right] + c_{t,1}^2 \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{\left(\varepsilon_{t,i}^2 > \frac{n\delta^2 \pi_t^{(n)}}{c_{t,1}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1 \right]$$

$$\leq \sum_{t=1}^{T} \max_{i \in [1:\, n]} \left\{ c_{t,0}^2 \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{\left(\varepsilon_{t,i}^2 > \frac{n\delta^2(1-\pi_t)}{c_{t,0}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 0 \right] + c_{t,1}^2 \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{\left(\varepsilon_{t,i}^2 > \frac{n\delta^2 \pi_t^{(n)}}{c_{t,1}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1 \right] \right\}$$

Note that for any $t \in [1:\, T]$ and $i \in [1:\, n]$,

$$\mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{\left(\varepsilon_{t,i}^2 > \frac{n\delta^2 \pi_t^{(n)}}{c_{t,1}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1 \right]$$

$$= \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{\left(\varepsilon_{t,i}^2 > \frac{n\delta^2 \pi_t^{(n)}}{c_{t,1}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1 \right] \left(\mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} + \mathbb{I}_{(\pi_t^{(n)} \notin [f(n), 1-f(n)])} \right)$$

$$\leq \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{\left(\varepsilon_{t,i}^2 > \frac{n\delta^2 f(n)}{c_{t,1}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1 \right] + \sigma^2 \mathbb{I}_{(\pi_t^{(n)} \notin [f(n), 1-f(n)])}$$

The second term converges in probability to zero as $n \to \infty$ by our clipping assumption. We now show how the first term goes to zero in probability. Since we assume $f(n) = \omega(\frac{1}{n})$, $nf(n) \to \infty$. So, it is sufficient to show that for all t, n,

$$\lim_{m \to \infty} \max_{t \in [1:\, n]} \left\{ \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(\varepsilon_{t,i}^2 > m)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1 \right] \right\} = 0$$

By Condition A.2.1, we have that for all $n \geq 1$,

$$\max_{t \in [1:\, T], i \in [1:\, n]} \mathbb{E}[\phi(\varepsilon_{t,i}^2) | H_{t-1}^{(n)}, A_{t,i} = 1] < M$$

Since we assume that $\lim_{x \to \infty} \frac{\phi(x)}{x} = \infty$, for all m, there exists a b_m s.t. $\phi(x) \geq mMx$ for all

$x \geq b_m$. So, for all n, t, i,

$$M \geq \mathbb{E}[\phi(\varepsilon_{t,i}^2)|H_{t-1}^{(n)}, A_{t,i} = 1] \geq \mathbb{E}[\phi(\varepsilon_{t,i}^2)\mathbb{I}_{(\varepsilon_{t,i}^2 \geq b_m)}|H_{t-1}^{(n)}, A_{t,i} = 1]$$

$$\geq mM\mathbb{E}[\varepsilon_{t,i}^2\mathbb{I}_{(\varepsilon_{t,i}^2 \geq b_m)}|H_{t-1}^{(n)}, A_{t,i} = 1]$$

Thus,

$$\max_{t\in[1:\,T],i\in[1:\,n]} \mathbb{E}[\varepsilon_{t,i}^2\mathbb{I}_{(\varepsilon_{t,i}^2 \geq b_m)}|H_{t-1}^{(n)}, A_{t,i} = 1] \leq \frac{1}{m}$$

We can make a very similar argument that for all $t \in [1:\,T]$, as $n \to \infty$,

$$\max_{i\in[1:\,n]} \mathbb{E}\left[\varepsilon_{t,i}^2\mathbb{I}_{\left(\varepsilon_{t,i}^2 > \frac{n\delta^2(1-\pi_t)}{c_{t,0}^2}\right)}\middle| H_{t-1}^{(n)}, A_{t,i} = 0\right] \xrightarrow{P} 0 \qquad \square$$

Corollary A.4.1 (Asymptotic Normality of the Batched OLS Estimator of Margin; 2-arm bandit setting)**.** *Assume the same conditions as Theorem 3.5.1. For each $t \in [1:\,T]$, we have the BOLS estimator of the margin $\beta_1 - \beta_0$:*

$$\hat{\Delta}_t^{\text{BOLS}} = \frac{\sum_{i=1}^n (1 - A_{t,i})R_{t,i}}{N_{t,0}} - \frac{\sum_{i=1}^n A_{t,i}R_{t,i}}{N_{t,1}}$$

We show that as $n \to \infty$,

$$\begin{bmatrix} \sqrt{\frac{N_{1,0}N_{1,1}}{n}}(\hat{\Delta}_1^{\text{BOLS}} - \Delta_1) \\ \sqrt{\frac{N_{2,0}N_{2,1}}{n}}(\hat{\Delta}_2^{\text{BOLS}} - \Delta_2) \\ \vdots \\ \sqrt{\frac{N_{T,0}N_{T,1}}{n}}(\hat{\Delta}_T^{\text{BOLS}} - \Delta_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_T)$$

PROOF:

$$\sqrt{\frac{N_{t,0}N_{t,1}}{n}}(\hat{\Delta}_t^{\text{BOLS}} - \Delta_t) = \sqrt{\frac{N_{t,0}N_{t,1}}{n}}\left(\frac{\sum_{i=1}^n (1-A_{t,i})\varepsilon_{t,i}}{N_{t,0}} - \frac{\sum_{i=1}^n A_{t,i}\varepsilon_{t,i}}{N_{t,1}}\right)$$

$$= \sqrt{\frac{N_{t,1}}{n}}\frac{\sum_{i=1}^n (1-A_{t,i})\varepsilon_{t,i}}{\sqrt{N_{t,0}}} - \sqrt{\frac{N_{t,0}}{n}}\frac{\sum_{i=1}^n A_{t,i}\varepsilon_{t,i}}{\sqrt{N_{t,1}}}$$

$$= \begin{bmatrix} \sqrt{\frac{N_{t,1}}{n}} & -\sqrt{\frac{N_{t,0}}{n}} \end{bmatrix} \begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i}$$

By Slutsky's Theorem and Lemma A.4.1, it is sufficient to show that as $n \to \infty$,

$$\begin{bmatrix} \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_1^{(n)}} & -\sqrt{1-\pi_1^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_1^{(n)} & 0 \\ 0 & \pi_1^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \varepsilon_{1,i} \\ \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_2^{(n)}} & -\sqrt{1-\pi_2^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_2^{(n)} & 0 \\ 0 & \pi_2^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \varepsilon_{2,i} \\ \vdots \\ \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_t^{(n)}} & -\sqrt{1-\pi_t^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \varepsilon_{T,i} \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_T)$$

By Cramer-Wold device, it is sufficient to show that for any fixed vector $\mathbf{d} \in \mathbb{R}^T$ s.t. $\|\mathbf{d}\|_2 =$

1 that

$$\mathbf{d}^\top \begin{bmatrix} \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_1^{(n)}} & -\sqrt{1-\pi_1^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_1^{(n)} & 0 \\ 0 & \pi_1^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \varepsilon_{1,i} \\ \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_2^{(n)}} & -\sqrt{1-\pi_2^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_2^{(n)} & 0 \\ 0 & \pi_2^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \varepsilon_{2,i} \\ \vdots \\ \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_t^{(n)}} & -\sqrt{1-\pi_t^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \varepsilon_{T,i} \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Let $[d_1, d_2, ..., d_T]^\top := \mathbf{d} \in \mathbb{R}^T$. The above is equivalent to

$$\sum_{t=1}^T \frac{1}{\sqrt{n}} d_t \begin{bmatrix} \sqrt{\pi_t^{(n)}} & -\sqrt{1-\pi_t^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Define $Y_{t,i} := \frac{1}{\sqrt{n}} d_t \begin{bmatrix} \sqrt{\pi_t^{(n)}} & -\sqrt{1-\pi_t^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i}$.

$\{Y_{1,1}, Y_{1,2}, ..., Y_{1,n}, ..., Y_{T,1}, Y_{T,2}, ..., Y_{T,n}\}$ is a martingale difference array with respect to the sequence of histories $\{H_t^{(n)}\}_{t=1}^T$ because for all $i \in [1\colon n]$ and $t \in [1\colon T]$,

$$\mathbb{E}[Y_{t,i}|H_{t-1}^{(n)}] = \frac{1}{\sqrt{n}} d_t \begin{bmatrix} \sqrt{\pi_t^{(n)}} & -\sqrt{1-\pi_t^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-1/2} \mathbb{E}\left[\begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \varepsilon_{t,i} \middle| H_{t-1}^{(n)} \right]$$

$$= \frac{d_t}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_t^{(n)}} & -\sqrt{1-\pi_t^{(n)}} \end{bmatrix} \begin{bmatrix} 1-\pi_t^{(n)} & 0 \\ 0 & \pi_t^{(n)} \end{bmatrix}^{-\frac{1}{2}} \mathbb{E}\left[\begin{bmatrix} (1-\pi_t^{(n)})\mathbb{E}[\varepsilon_{t,i}|H_{t-1}^{(n)}, A_{t,i}=0] \\ \pi_{t,i}\mathbb{E}[\varepsilon_{t,i}|H_{t-1}^{(n)}, A_{t,i}=1] \end{bmatrix} \middle| H_{t-1}^{(n)} \right] = 0$$

We now apply[29] martingale central limit theorem to $Y_{t,i}$ to show the desired result. Verify-

ing the conditions for the martingale CLT is equivalent to what we did to verify the conditions in the conditions in the proof of Theorem 3.5.1—the only difference is that we replace $\mathbf{c}_t^\top$ in the Theorem 3.5.1 proof with $d_t \begin{bmatrix} \sqrt{1-\pi_t^{(n)}} & -\sqrt{\pi_t^{(n)}} \end{bmatrix}$ in this proof. Even though $\mathbf{c}_t$ is a constant vector and $d_t \begin{bmatrix} \sqrt{1-\pi_t^{(n)}} & -\sqrt{\pi_t^{(n)}} \end{bmatrix}$ is a random vector, the proof still goes through with this adjusted $\mathbf{c}_t$ vector, since (i) $d_t \begin{bmatrix} \sqrt{1-\pi_t^{(n)}} & -\sqrt{\pi_t^{(n)}} \end{bmatrix} \in H_{t-1}^{(n)}$, (ii) $\| \begin{bmatrix} \sqrt{1-\pi_t^{(n)}} & -\sqrt{\pi_t^{(n)}} \end{bmatrix} \|_2 = 1$, and (iii) $\frac{n\delta^2 \pi_t^{(n)}}{c_{t,1}^2} = \frac{n\delta^2 \pi_t^{(n)}}{d_t^2 \pi_t^{(n)}} \to \infty$ and $\frac{n\delta^2(1-\pi_t)}{c_{t,0}^2} = \frac{n\delta^2(1-\pi_t)}{d_t^2(1-\pi_t)} \to \infty$. $\square$

Corollary A.4.2 (Consistency of BOLS Variance Estimator)**.** *Assuming Conditions 3.4.1 (moments) and 3.4.3 (conditionally i.i.d. actions), and a clipping rate of $f(n) = \omega(\frac{1}{n})$ (Definition 3.3.1), for all $t \in [1\colon T]$, as $n \to \infty$,*

$$\hat{\sigma}_t^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(R_{t,i} - A_{t,i}\hat{\beta}_{t,1}^{\text{BOLS}} - (1-A_{t,i})\hat{\beta}_{t,0}^{\text{BOLS}}\right)^2 \xrightarrow{P} \sigma^2$$

Proof:

$$\hat{\sigma}_t^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(R_{t,i} - A_{t,i}\hat{\beta}_{t,1}^{\text{BOLS}} - (1-A_{t,i})\hat{\beta}_{t,0}^{\text{BOLS}}\right)^2$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}\left(\left[A_{t,i}\beta_{t,1} + (1-A_{t,i})\beta_{t,0} + \varepsilon_{t,i}\right] - A_{t,i}\left[\beta_{t,1} + \frac{\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}}{N_{t,1}}\right] - (1-A_{t,i})\left[\beta_{t,0} + \frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{N_{t,0}}\right]\right)^2$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}\left(\varepsilon_{t,i} - A_{t,i}\frac{\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}}{N_{t,1}} - (1-A_{t,i})\frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{N_{t,0}}\right)^2$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}\left(\varepsilon_{t,i}^2 - 2A_{t,i}\varepsilon_{t,i}\frac{\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}}{N_{t,1}} - 2(1-A_{t,i})\varepsilon_{t,i}\frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{N_{t,0}} + A_{t,i}\left[\frac{\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}}{N_{t,1}}\right]^2 + (1-A_{t,i})\left[\frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{N_{t,0}}\right]^2\right)$$

$$= \left(\frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_{t,i}^2\right) - 2\frac{(\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i})^2}{(n-2)N_{t,1}} - 2\frac{(\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i})^2}{(n-2)N_{t,0}} + \frac{N_{t,1}}{n-2}\left[\frac{\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}}{N_{t,1}}\right]^2 + \frac{N_{t,0}}{n-2}\left[\frac{\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i}}{N_{t,0}}\right]^2$$

$$= \left(\frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_{t,i}^2\right) - \frac{(\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i})^2}{(n-2)N_{t,1}} - \frac{(\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i})^2}{(n-2)N_{t,0}}$$

Note that $\frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_{t,i}^2 \xrightarrow{P} \sigma^2$ because for all $\delta > 0$,

$$\mathbb{P}\left(\left|\left[\frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_{t,i}^2\right] - \sigma^2\right| > \delta\right)$$

$$\leq \mathbb{P}\left(\left|\left[\frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_{t,i}^2\right] - \frac{\sigma^2(n-2)}{n}\right| > \delta/2\right) + \mathbb{P}\left(\left|\frac{\sigma^2(n-2)}{n} - \sigma^2\right| > \delta/2\right)$$

$$= \mathbb{P}\left(\left|\frac{1}{n-2}\sum_{i=1}^{n}(\varepsilon_{t,i}^2 - \sigma^2)\right| > \delta/2\right) + \mathbb{P}\left(\sigma^2\left|\frac{-2}{n}\right| > \delta/2\right)$$

Since the second term in the summation above goes to zero for sufficiently large n, we now focus on the first term in the summation above. By Chebychev inequality,

$$\mathbb{P}\left(\left|\frac{1}{n-2}\sum_{i=1}^{n}(\varepsilon_{t,i}^2 - \sigma^2)\right| > \delta/2\right) \leq \frac{4}{\delta^2(n-2)^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(\varepsilon_{t,i}^2 - \sigma^2)(\varepsilon_{t,j}^2 - \sigma^2)\right]$$

$$= \frac{4}{\delta^2(n-2)^2}\mathbb{E}\left[\sum_{i=1}^{n}(\varepsilon_{t,i}^2 - \sigma^2)^2\right]$$

where the equality above holds because for $i \neq j$, $\mathbb{E}[(\varepsilon_{t,i}^2 - \sigma^2)(\varepsilon_{t,j}^2 - \sigma^2)] = \mathbb{E}\big[\mathbb{E}[(\varepsilon_{t,i}^2 - \sigma^2)(\varepsilon_{t,j}^2 - \sigma^2)|\mathcal{H}_{t-1}^{(n)}]\big] = \mathbb{E}\big[\mathbb{E}[\varepsilon_{t,i}^2 - \sigma^2|\mathcal{H}_{t-1}^{(n)}]\mathbb{E}[\varepsilon_{t,j}^2 - \sigma^2|\mathcal{H}_{t-1}^{(n)}]\big] = 0$. By Condition 3.4.1

$\mathbb{E}[\varepsilon_{t,i}^4|\mathcal{H}_{t=1}^{(n)}] < M < \infty$,

$$= \frac{4}{\delta^2(n-2)^2}\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}[(\varepsilon_{t,i}^4 - 2\varepsilon_{t,i}^2\sigma^2 + \sigma^4)|\mathcal{H}_{t-1}^{(n)}]\right] \leq \frac{4n(M+\sigma^4)}{\delta^2(n-2)^2} \to 0$$

Thus by Slutsky's Theorem it is sufficient to show that $\frac{(\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i})^2}{(n-2)N_{t,1}} + \frac{(\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i})^2}{(n-2)N_{t,0}} \xrightarrow{P} 0$. We will only show that $\frac{(\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i})^2}{(n-2)N_{t,1}} \xrightarrow{P} 0$; $\frac{(\sum_{i=1}^{n}(1-A_{t,i})\varepsilon_{t,i})^2}{(n-2)N_{t,0}} \xrightarrow{P} 0$ holds by a very similar argument.

Note that by Lemma A.4.1, $\frac{N_{t,1}}{n\pi_t^{(n)}} \xrightarrow{P} 1$. Thus, to show that $\frac{(\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i})^2}{(n-2)N_{t,1}} \xrightarrow{P} 0$ by Slutsky's Theorem it is sufficient to show that $\frac{(\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i})^2}{(n-2)n\pi_t^{(n)}} \xrightarrow{P} 0$. Let $\delta > 0$. By Markov inequality,

$$\mathbb{P}\left(\left|\frac{(\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i})^2}{(n-2)n\pi_t^{(n)}}\right| > \delta\right) \leq \mathbb{E}\left[\frac{1}{\delta(n-2)n\pi_t^{(n)}}\left(\sum_{i=1}^{n}A_{t,i}\varepsilon_{t,i}\right)^2\right]$$

$$= \mathbb{E}\left[\frac{1}{\delta(n-2)n\pi_t^{(n)}}\sum_{j=1}^{n}\sum_{i=1}^{n}A_{t,j}A_{t,i}\varepsilon_{t,i}\varepsilon_{t,j}\right]$$

Since $\pi_t^{(n)} \in \mathcal{H}_{t-1}^{(n)}$,

$$= \mathbb{E}\left[\frac{1}{\delta(n-2)n\pi_t^{(n)}}\sum_{j=1}^{n}\sum_{i=1}^{n}\mathbb{E}[A_{t,j}A_{t,i}\varepsilon_{t,i}\varepsilon_{t,j}|\mathcal{H}_{t-1}^{(n)}]\right]$$

Since for $i \neq j$, $\mathbb{E}[A_{t,j}A_{t,i}\varepsilon_{t,j}\varepsilon_{t,i}|\mathcal{H}_{t-1}^{(n)}] = \mathbb{E}[A_{t,j}\varepsilon_{t,j}|\mathcal{H}_{t-1}^{(n)}]\mathbb{E}[A_{t,i}\varepsilon_{t,i}|\mathcal{H}_{t-1}^{(n)}] = 0$,

$$= \mathbb{E}\left[\frac{1}{\delta(n-2)n\pi_t^{(n)}}\sum_{i=1}^{n}\mathbb{E}[A_{t,i}\varepsilon_{t,i}^2|\mathcal{H}_{t-1}^{(n)}]\right]$$

Since $\mathbb{E}[A_{t,i}\varepsilon_{t,i}^2|\mathcal{H}_{t-1}^{(n)}] = \mathbb{E}[\varepsilon_{t,i}^2|\mathcal{H}_{t-1}^{(n)}, A_{t,i}=1]\pi_t^{(n)} = \sigma^2\pi_t^{(n)}$,

$$= \mathbb{E}\left[\frac{1}{\delta(n-2)n\pi_t^{(n)}} n\sigma^2\pi_t^{(n)}\right] = \frac{\sigma^2}{\delta(n-2)} \to 0 \qquad \square$$

A.5 Asymptotic Normality of the Batched OLS Estimator: Contextual Bandits

Theorem 3.5.2 (Asymptotic Normality of the Batched OLS Statistic) *For a K-armed contextual bandit, we for each $t \in [1:T]$, we have the BOLS estimator:*

$$\hat{\beta}_t^{\text{BOLS}} = \begin{bmatrix} \underline{\boldsymbol{C}}_{t,0} & 0 & 0 & \dots & 0 \\ 0 & \underline{\boldsymbol{C}}_{t,1} & 0 & \dots & 0 \\ 0 & 0 & \underline{\boldsymbol{C}}_{t,2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \underline{\boldsymbol{C}}_{t,K-1} \end{bmatrix}^{-1} \sum_{i=1}^{n} \begin{bmatrix} \mathbb{I}_{A_{t,i}=0} \boldsymbol{C}_{t,i} \\ \mathbb{I}_{A_{t,i}=1} \boldsymbol{C}_{t,i} \\ \vdots \\ \mathbb{I}_{A_{t,i}=K-1} \boldsymbol{C}_{t,i} \end{bmatrix} R_{t,i} \in \mathbb{R}^{Kd}$$

where $\underline{\boldsymbol{C}}_{t,k} \triangleq \sum_{i=1}^{n} \mathbb{I}_{A_{t,i}^{(n)}=k} \boldsymbol{C}_{t,i} (\boldsymbol{C}_{t,i})^\top \in \mathbb{R}^{d \times d}$. Assuming Conditions A.2.1 (weak moments), 3.4.3 (conditionally i.i.d. actions), 3.5.1 (conditionally i.i.d. contexts), and 3.5.2 (bounded contexts), and a conditional clipping rate $f(n) = c$ for some $0 \leq c < \frac{1}{2}$ (see Definition 3.5.1), we show that as $n \to \infty$,

$$\begin{bmatrix} \text{Diagonal}\left[\underline{\boldsymbol{C}}_{1,0}, \underline{\boldsymbol{C}}_{1,1}, ..., \underline{\boldsymbol{C}}_{1,K-1}\right]^{1/2} (\hat{\beta}_1^{\text{BOLS}} - \beta_1) \\ \text{Diagonal}\left[\underline{\boldsymbol{C}}_{2,0}, \underline{\boldsymbol{C}}_{2,1}, ..., \underline{\boldsymbol{C}}_{2,K-1}\right]^{1/2} (\hat{\beta}_2^{\text{BOLS}} - \beta_2) \\ \vdots \\ \text{Diagonal}\left[\underline{\boldsymbol{C}}_{T,0}, \underline{\boldsymbol{C}}_{T,1}, ..., \underline{\boldsymbol{C}}_{T,K-1}\right]^{1/2} (\hat{\beta}_T^{\text{BOLS}} - \beta_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_{TKd})$$

Lemma A.5.1. *Assuming the conditions of Theorem 3.5.2, for any batch $t \in [1:T]$ and any*

arm $k \in [0 : K-1]$, *as* $n \to \infty$,

$$\left[\sum_{i=1}^{n} \mathbb{I}_{A_{t,i}=k} \boldsymbol{C}_{t,i} \boldsymbol{C}_{t,i}^{\top}\right] \left[n \underline{\boldsymbol{Z}}_{t,k} P_{t,k}\right]^{-1} \xrightarrow{P} \underline{\boldsymbol{I}}_{d} \tag{A.5.1}$$

$$\left[\sum_{i=1}^{n} \mathbb{I}_{A_{t,i}=k} \boldsymbol{C}_{t,i} \boldsymbol{C}_{t,i}^{\top}\right]^{1/2} \left[n \underline{\boldsymbol{Z}}_{t,k} P_{t,k}\right]^{-1/2} \xrightarrow{P} \underline{\boldsymbol{I}}_{d} \tag{A.5.2}$$

where $P_{t,k} \triangleq \mathbb{P}(A_{t,i} = k | H_{t-1}^{(n)})$ *and* $\underline{\boldsymbol{Z}}_{t,k} \triangleq \mathbb{E}\left[\boldsymbol{C}_{t,i} \boldsymbol{C}_{t,i}^{\top} \middle| H_{t-1}^{(n)}, A_{t,i} = k\right]$.

Proof of Lemma A.5.1: We first show that as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top} - \underline{\mathbf{Z}}_{t,k}P_{t,k}\right) \xrightarrow{P} \underline{0}$. It is sufficient to show that convergence holds entry-wise so for any $r, s \in [0 : d-1]$, as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s) - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right) \xrightarrow{P} 0$. Note that

$$\mathbb{E}\left[\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s) - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s)\middle| H_{t-1}, A_{t,i} = k\right]P_{t,k} - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right] = 0$$

By Chebychev inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s) - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right| > \varepsilon\right)$$

$$\leq \frac{1}{\varepsilon^2 n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top} - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right)^2\right]$$

$$= \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\left[\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s) - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right]\left[\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,j}\mathbf{C}_{t,j}^{\top}(r,s) - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right]\right] \tag{A.5.3}$$

By conditional independence and by law of iterated expectations (conditioning on $H_{t-1}^{(n)}$), for $i \neq j$, $\mathbb{E}\left[\left(\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s) - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right)\left(\mathbb{I}_{A_{t,j}=k}\mathbf{C}_{t,j}\mathbf{C}_{t,j}^{\top}(r,s) - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right)\right] = 0$. Thus, (A.5.3) above equals the following:

$$= \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s) - P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s)\right)^2\right]$$

$$= \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{I}_{A_{t,i}=k}\left[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s)\right]^2 - 2\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s)P_{t,k}\underline{\mathbf{Z}}_{t,k}(r,s) + P_{t,k}^2\left[\underline{\mathbf{Z}}_{t,k}(r,s)\right]^2\right]$$

$$= \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{I}_{A_{t,i}=k}\left[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s)\right]^2 - P_{t,k}^2\left[\underline{\mathbf{Z}}_{t,k}(r,s)\right]^2\right]$$

$$= \frac{1}{\varepsilon^2 n}\mathbb{E}\left[\mathbb{I}_{A_{t,i}=k}\left[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}(r,s)\right]^2 - P_{t,k}^2\left[\underline{\mathbf{Z}}_{t,k}(r,s)\right]^2\right] \leq \frac{2d\max(u^2,1)}{\varepsilon^2 n} \to 0$$

as $n \to \infty$. The last inequality above holds by Condition 3.5.2.

Proving Equation (A.5.1)**:**

It is sufficient to show that

$$\left\|\frac{2\max(du^2,1)}{\varepsilon^2 n}\left[n\underline{\mathbf{Z}}_{t,k}P_{t,k}\right]^{-1}\right\|_{op} = \left\|\frac{2\max(du^2,1)}{\varepsilon^2 n^2 P_{t,k}}\underline{\mathbf{Z}}_{t,k}^{-1}\right\|_{op} \xrightarrow{P} 0 \tag{A.5.4}$$

We define random variable $M_t^{(n)} = \mathbb{I}_{(\forall\, \mathbf{c}\in\mathbb{R}^d,\, \mathcal{A}_t(H_{t-1}^{(n)},\mathbf{c})\in[f(n),1-f(n)]^K)}$, representing whether the conditional clipping condition is satisfied. Note that by our conditional clipping assumption, $M_t^{(n)} \xrightarrow{P} 1$ as $n \to \infty$. The left hand side of (A.5.4) is equal to the following

$$\left\|\frac{2\max(du^2,1)}{\varepsilon^2 n^2 P_{t,k}}\underline{\mathbf{Z}}_{t,k}^{-1}(M_t^{(n)} + (1-M_t^{(n)}))\right\|_{op} = \left\|\frac{2\max(du^2,1)}{\varepsilon^2 n^2 P_{t,k}}\underline{\mathbf{Z}}_{t,k}^{-1}M_t^{(n)}\right\|_{op} + o_p(1) \tag{A.5.5}$$

By our conditional clipping condition and Bayes rule we have that for all $\mathbf{c} \in [-u, u]^d$,

$$\mathbb{P}(\mathbf{C}_{t,i} = \mathbf{c} | A_{t,i} = k, H_{t-1}^{(n)}, M_t^{(n)} = 1)$$

$$= \frac{\mathbb{P}(A_{t,i} = k | \mathbf{C}_{t,i} = \mathbf{c}, H_{t-1}^{(n)}, M_t^{(n)} = 1) \mathbb{P}(\mathbf{C}_{t,i} = \mathbf{c} | H_{t-1}^{(n)}, M_t^{(n)} = 1)}{\mathbb{P}(A_{t,i} = k | H_{t-1}^{(n)}, M_t^{(n)} = 1)}$$

$$\geq \frac{f(n)\, \mathbb{P}(\mathbf{C}_{t,i} = \mathbf{c} | H_{t-1}^{(n)}, M_t^{(n)} = 1)}{1}.$$

Thus, we have that

$$\underline{\mathbf{Z}}_{t,k} M_t^{(n)} = \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)}, A_{t,i} = k\big] M_t^{(n)} = \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)}, A_{t,i} = k, M_t^{(n)} = 1\big] M_t^{(n)}$$

$$\succcurlyeq f(n) \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)}, M_t^{(n)} = 1\big] M_t^{(n)} = f(n) \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)}\big] M_t^{(n)} = f(n) \underline{\Sigma}_t^{(n)} M_t^{(n)}.$$

By apply matrix inverses to both sides of the above inequality, we get that

$$\lambda_{\max}(\underline{\mathbf{Z}}_{t,k}^{-1} M_t^{(n)}) \leq \frac{1}{f(n)} \lambda_{\max}\left(\left(\underline{\Sigma}_t^{(n)}\right)^{-1}\right) M_t^{(n)} \leq \frac{1}{l f(n)} \tag{A.5.6}$$

where the last inequality above holds for constant l by Condition 3.5.2. Recall that $P_{t,k} = \mathbb{P}(A_{t,i} = k \mid H_{t-1}^{(n)})$, so $P_{t,k} \mid (M_t^{(n)} = 1) \geq f(n)$. Thus, equation (A.5.5) is bounded above by the following

$$\leq \frac{2 \max(du^2, 1)}{\varepsilon^2 n^2 l f(n)^2} + o_p(1) \xrightarrow{P} 0$$

where the limit above holds because we assume that $f(n) = c$ for some $0 < c \leq \frac{1}{2}$ $\square$.

Proving Equation (A.5.2)**:** By Condition 3.5.2, $\|\frac{1}{n}\underline{\mathbf{C}}_{t,k}\|_{\max} \leq u$ and $\|\underline{\mathbf{Z}}_{t,k} P_{t,k}\|_{\max} \leq u$. Thus, any continuous function of $\frac{1}{n}\underline{\mathbf{C}}_{t,k}$ and $\underline{\mathbf{Z}}_{t,k} P_{t,k}$ will have compact support and thus

be uniformly continuous. For any uniformly continuous function $f : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that for any matrices $\underline{\mathbf{A}}, \underline{\mathbf{B}} \in \mathbb{R}^{d\times d}$, whenever $\|\underline{\mathbf{A}} - \underline{\mathbf{B}}\|_{\text{op}} < \delta$, then $\|f(\underline{\mathbf{A}}) - f(\underline{\mathbf{B}})\|_{\text{op}} < \varepsilon$. Thus, for any $\varepsilon > 0$, there exists some $\delta > 0$ such that

$$\mathbb{P}\left(\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(A_{t,k}=k)}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}\right) - \underline{\mathbf{Z}}_{t,k}P_{t,k}\right\|_{\text{op}} > \delta\right) \to 0$$

implies

$$\mathbb{P}\left(\left\|f\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(A_{t,k}=k)}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}\right) - f(\underline{\mathbf{Z}}_{t,k}P_{t,k})\right\|_{\text{op}} > \varepsilon\right) \to 0$$

Thus, by letting f be the matrix square-root function,

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(A_{t,k}=k)}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}\right)^{1/2} - (\underline{\mathbf{Z}}_{t,k}P_{t,k})^{1/2} \xrightarrow{P} \underline{0}.$$

We now want to show that for some constant $r > 0$, $\mathbb{P}\left(\left\|\underline{\mathbf{Z}}_{t,k}^{-1}\frac{1}{P_{t,k}}\right\|_{\text{op}} > r\right)$, because this would imply that

$$\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(A_{t,k}=k)}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^{\top}\right)^{1/2} - (\underline{\mathbf{Z}}_{t,k}P_{t,k})^{1/2}\right](\underline{\mathbf{Z}}_{t,k}P_{t,k})^{-1/2} \xrightarrow{P} \underline{0}.$$

Recall that for $M_t^{(n)} = \mathbb{I}_{(\forall\, \mathbf{c}\in\mathbb{R}^d,\, \mathcal{A}_t(H_{t-1}^{(n)},\mathbf{c})\in[f(n),1-f(n)]^K)}$, representing whether the conditional clipping condition is satisfied,

$$\underline{\mathbf{Z}}_{t,k}^{-1} = \underline{\mathbf{Z}}_{t,k}^{-1}(M_t^{(n)} + (1 - M_t^{(n)})) = \underline{\mathbf{Z}}_{t,k}^{-1}M_t^{(n)} + o_p(1).$$

Thus it is sufficient to show that $\mathbb{P}\left(\left\|\underline{\mathbf{Z}}_{t,k}^{-1}\frac{1}{P_{t,k}}M_t^{(n)}\right\|_{\text{op}} > r\right)$. Recall that by equation (A.5.6)

we have that

$$\lambda_{\max}(\underline{\mathbf{Z}}_{t,k}^{-1}\mathcal{M}_t^{(n)}) \leq \frac{1}{f(n)}\lambda_{\max}\left(\left(\underline{\Sigma}_t^{(n)}\right)^{-1}\right)\mathcal{M}_t^{(n)} \leq \frac{1}{lf(n)}$$

Also note that $P_{t,k} = \mathbb{P}(\mathcal{A}_{t,i} = k \mid H_{t-1}^{(n)})$, so $P_{t,k} \mid (\mathcal{M}_t^{(n)} = 1) \geq f(n)$. Thus we have that

$$\mathbb{P}\left(\left\|\underline{\mathbf{Z}}_{t,k}^{-1}\frac{1}{P_{t,k}}\mathcal{M}_t^{(n)}\right\|_{\text{op}} > r\right) \leq \mathbb{I}_{(\frac{1}{lf(n)^2} > r)} = 0$$

for $r > \frac{1}{lf(n)^2} = \frac{1}{lc^2}$, since we assume that $f(n) = c$ for some $0 < c \leq \frac{1}{2}$. $\square$

Proof of Theorem 3.5.2: Let $\underline{\mathbf{Z}}_{t,k} \triangleq \mathbb{E}\left[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \middle| H_{t-1}^{(n)}, \mathcal{A}_{t,i} = k\right]$ and $P_{t,k} \triangleq \mathbb{P}(\mathcal{A}_{t,i} = k | H_{t-1}^{(n)})$. We also define

$$\mathbf{D}_t^{(n)} \triangleq \text{Diagonal}\left[\underline{\mathbf{C}}_{t,0}, \underline{\mathbf{C}}_{t,1}, ..., \underline{\mathbf{C}}_{t,K-1}\right]^{1/2}(\hat{\beta}_t - \beta_t) = \sum_{i=1}^{n}\begin{bmatrix} \underline{\mathbf{C}}_{t,0}^{-1/2}\mathbf{C}_{t,i}\mathbb{I}_{\mathcal{A}_{t,i}=0} \\ \underline{\mathbf{C}}_{t,1}^{-1/2}\mathbf{C}_{t,i}\mathbb{I}_{\mathcal{A}_{t,i}=1} \\ \vdots \\ \underline{\mathbf{C}}_{t,K-1}^{-1/2}\mathbf{C}_{t,i}\mathbb{I}_{\mathcal{A}_{t,i}=K-1} \end{bmatrix}\varepsilon_{t,i}$$

We want to show that $[\mathbf{D}_1^{(n)}, \mathbf{D}_2^{(n)}, ..., \mathbf{D}_T^{(n)}]^\top \xrightarrow{D} \mathcal{N}(0, \sigma^2\underline{\mathbf{I}}_{TKd})$. By Lemma A.5.1 and Slutsky's Theorem, it sufficient to show that as $n \to \infty$, $[\mathbf{Q}_1^{(n)}, \mathbf{Q}_2^{(n)}, ..., \mathbf{Q}_T^{(n)}]^\top \xrightarrow{D}$

$\mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_{TKd})$ for

$$\mathbf{Q}_t^{(n)} \triangleq \sum_{i=1}^{n} \begin{bmatrix} \frac{1}{\sqrt{nP_{t,0}}} \underline{\mathbf{Z}}_{t,0}^{-1/2} \mathbf{C}_{t,i} \mathbb{I}_{A_{t,i}=0} \\ \frac{1}{\sqrt{nP_{t,1}}} \underline{\mathbf{Z}}_{t,1}^{-1/2} \mathbf{C}_{t,i} \mathbb{I}_{A_{t,i}=1} \\ \vdots \\ \frac{1}{\sqrt{nP_{t,K-1}}} \underline{\mathbf{Z}}_{t,K-1}^{-1/2} \mathbf{C}_{t,i} \mathbb{I}_{A_{t,i}=K-1} \end{bmatrix} \varepsilon_{t,i}$$

By Cramer Wold device, it is sufficient to show that for any $\mathbf{b} \in \mathbb{R}^{TKd}$ with $\|\mathbf{b}\|_2 = 1$, where $\mathbf{b} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_T]$ for $\mathbf{b}_t \in \mathbb{R}^{Kd}$, as $n \to \infty$.

$$\sum_{t=1}^{T} \mathbf{b}_t^\top \mathbf{Q}_t^{(n)} \xrightarrow{D} \mathcal{N}(0, \sigma^2) \tag{A.5.7}$$

We can further define for all $t \in [1\colon T]$, $\mathbf{b}_t = [\mathbf{b}_{t,0}, \mathbf{b}_{t,1}, ..., \mathbf{b}_{t,K-1}]$ with $\mathbf{b}_{t,k} \in \mathbb{R}^d$. Thus to show (A.5.7) it is equivalent to show that

$$\sum_{t=1}^{T} \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \frac{1}{\sqrt{nP_{t,k}}} \underline{\mathbf{Z}}_{t,k}^{-1/2} \sum_{i=1}^{n} \mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \varepsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

We define $Y_{t,i}^{(n)} \triangleq \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \frac{1}{\sqrt{nP_{t,k}}} \mathbb{I}_{A_{t,i}=k} \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{C}_{t,i} \varepsilon_{t,i}$. The sequence $Y_{1,1}^{(n)}$, $Y_{1,2}^{(n)}$, ..., $Y_{1,n}^{(n)}$, ... $Y_{T,1}^{(n)}$, $Y_{T,2}^{(n)}$, ..., $Y_{T,n}^{(n)}$ is a martingale difference array with respect to the sequence of histories $\{H_{t-1}^{(n)}\}_{t=1}^{T}$ because $\mathbb{E}[Y_{t,i}^{(n)} | H_{t-1}^{(n)}] = \mathbb{E}\left[\mathbb{E}[Y_{t,i}^{(n)} | H_{t-1}^{(n)}, A_{t,i}, \mathbf{C}_{t,i}] \middle| H_{t-1}^{(n)}\right] = 0$ for all $i \in [1\colon n]$ and all $t \in [1\colon T]$. We then apply the martingale central limit theorem of[29] to $Y_{t,i}^{(n)}$ to show the desired result (see the proof of Theorem 3.4.1 in Appendix A.2 for the statement of the martingale CLT conditions). Note that the first condition (a) of the martingale CLT is already satisfied, as we just showed that $Y_{t,i}^{(n)}$ form a martingale difference array with respect to $H_{t-1}^{(n)}$.

Condition(b): Conditional Variance

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}[Y_{t,i}^2|H_{t-1}^{(n)}] = \sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\left[\left(\sum_{k=0}^{K-1}\mathbf{b}_{t,k}^\top\frac{1}{\sqrt{nP_{t,k}}}\underline{\mathbf{Z}}_{t,k}^{-1/2}\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\varepsilon_{t,i}\right)^2\middle|H_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\frac{1}{nP_{t,k}}\mathbf{b}_{t,k}^\top\underline{\mathbf{Z}}_{t,k}^{-1/2}\mathbb{E}\left[\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top\varepsilon_{t,i}^2\middle|H_{t-1}^{(n)}\right]\underline{\mathbf{Z}}_{t,k}^{-1/2}\mathbf{b}_{t,k}$$

By law of iterated expectations (conditioning on $H_{t-1}^{(n)}, A_{t,i}, \mathbf{C}_{t,i}$) and Condition A.2.1,

$$= \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\frac{1}{P_{t,k}}\mathbf{b}_{t,k}^\top\underline{\mathbf{Z}}_{t,k}^{-1/2}\mathbb{E}\left[\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top\middle|H_{t-1}^{(n)}\right]\underline{\mathbf{Z}}_{t,k}^{-1/2}\mathbf{b}_{t,k}\sigma^2$$

$$= \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\frac{1}{P_{t,k}}\mathbf{b}_{t,k}^\top\underline{\mathbf{Z}}_{t,k}^{-1/2}\mathbb{E}\left[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top\middle|H_{t-1}^{(n)}, A_{t,i}=k\right]P_{t,k}\underline{\mathbf{Z}}_{t,k}^{-1/2}\mathbf{b}_{t,k}\sigma^2$$

$$= \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbf{b}_{t,k}^\top\underline{\mathbf{I}}_d\mathbf{b}_{t,k}\sigma^2 = \sigma^2\sum_{t=1}^{T}\sum_{k=0}^{K-1}\mathbf{b}_{t,k}^\top\mathbf{b}_{t,k} = \sigma^2$$

Condition(c): Lindeberg Condition Let $\delta > 0$.

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\left[Y_{t,i}^2\mathbb{I}_{(|Y_{t,i}|>\delta)}\middle|H_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\left[\left(\sum_{k=0}^{K-1}\mathbf{b}_{t,k}^\top\frac{1}{\sqrt{nP_{t,k}}}\mathbf{Z}_{t,i}^{-1/2}\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\varepsilon_{t,i}\right)^2\mathbb{I}_{(Y_{t,i}^2>\delta^2)}\middle|H_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\frac{1}{nP_{t,k}}\mathbf{b}_{t,k}^\top\mathbf{Z}_{t,i}^{-1/2}\mathbb{E}\left[\mathbb{I}_{A_{t,i}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top\varepsilon_{t,i}^2\mathbb{I}_{(Y_{t,i}^2>\delta^2)}\middle|H_{t-1}^{(n)}\right]\mathbf{Z}_{t,i}^{-1/2}\mathbf{b}_{t,k}$$

It is sufficient to show that for any $t \in [1\colon T]$ and any $k \in [0 \colon K-1]$ the following converges in probability to zero:

$$\sum_{i=1}^{n} \frac{1}{nP_{t,k}} \mathbf{b}_{t,k}^{\top} \mathbf{Z}_{t,i}^{-1/2} \mathbb{E}\left[\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^{\top} \varepsilon_{t,i}^{2} \mathbb{I}_{(Y_{t,i}^{2} > \delta^{2})} \middle| H_{t-1}^{(n)} \right] \mathbf{Z}_{t,i}^{-1/2} \mathbf{b}_{t,k}$$

Recall that $Y_{t,i} = \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^{\top} \frac{1}{\sqrt{nP_{t,k}}} \mathbb{I}_{A_{t,i}=k} \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{C}_{t,i} \varepsilon_{t,i}$.

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbf{b}_{t,k}^{\top} \mathbf{Z}_{t,i}^{-1/2} \mathbb{E}\left[\mathbf{C}_{t,i} \mathbf{C}_{t,i}^{\top} \varepsilon_{t,i}^{2} \mathbb{I}_{(\frac{1}{nP_{t,k}} \mathbf{b}_{t,k}^{\top} \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^{\top} \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{b}_{t,k} \varepsilon_{t,i}^{2} > \delta^{2})} \middle| H_{t-1}^{(n)}, \mathcal{A}_{t,i} = k \right] \mathbf{Z}_{t,i}^{-1/2} \mathbf{b}_{t,k}$$

Since $\mathbf{c} \in [-u, u]$, by the Gershgorin circle theorem, we can bound the maximum eigenvalue of $\mathbf{c}\mathbf{c}^{\top}$ by some constant $a > 0$.

$$\leq \frac{a}{n} \sum_{i=1}^{n} \mathbf{b}_{t,k}^{\top} \mathbf{Z}_{t,i}^{-1} \mathbf{b}_{t,k} \mathbb{E}\left[\varepsilon_{t,i}^{2} \mathbb{I}_{(\frac{a}{nP_{t,k}} \mathbf{b}_{t,k}^{\top} \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \varepsilon_{t,i}^{2} > \delta^{2})} \;\middle|\; H_{t-1}^{(n)}, \mathcal{A}_{t,i} = k \right]$$

We define random variable $\mathcal{M}_{t}^{(n)} = \mathbb{I}_{(\forall\, \mathbf{c} \in \mathbb{R}^{d},\, \mathcal{A}_{t}(H_{t-1}^{(n)}, \mathbf{c}) \in [f(n), 1-f(n)]^{K})}$, representing whether the conditional clipping condition is satisfied. Note that by our conditional clipping assumption, $\mathcal{M}_{t}^{(n)} \xrightarrow{P} 1$ as $n \to \infty$.

$$= \frac{a}{n} \sum_{i=1}^{n} \mathbf{b}_{t,k}^{\top} \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \mathbb{E}\left[\varepsilon_{t,i}^{2} \mathbb{I}_{(\frac{a}{nP_{t,k}} \mathbf{b}_{t,k}^{\top} \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \varepsilon_{t,i}^{2} > \delta^{2})} \middle| H_{t-1}^{(n)}, \mathcal{A}_{t,i} = k \right] \left(\mathcal{M}_{t}^{(n)} + (1 - \mathcal{M}_{t}^{(n)}) \right)$$

$$= \frac{a}{n} \sum_{i=1}^{n} \mathbf{b}_{t,k}^{\top} \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \mathbb{E}\left[\varepsilon_{t,i}^{2} \mathbb{I}_{(\frac{a}{nP_{t,k}} \mathbf{b}_{t,k}^{\top} \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \varepsilon_{t,i}^{2} > \delta^{2})} \middle| H_{t-1}^{(n)}, \mathcal{A}_{t,i} = k \right] \mathcal{M}_{t}^{(n)} + o_{p}(1) \qquad \text{(A.5.8)}$$

By equation (A.5.6), have that

$$\lambda_{\max}(\underline{\mathbf{Z}}_{t,k}^{-1}) \leq \frac{1}{f(n)}\lambda_{\max}\left(\left(\underline{\Sigma}_t^{(n)}\right)^{-1}\right) \leq \frac{1}{lf(n)}$$

Recall that $P_{t,k} = \mathbb{P}(A_{t,i} = k \mid H_{t-1}^{(n)})$, so $P_{t,k} \mid (M_t^{(n)} = 1) \geq f(n)$. Thus we have that equation (A.5.8) is upper bounded by the following:

$$\leq \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}}{lf(n)} \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(\frac{a}{nf(n)} \frac{\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}}{lf(n)} \varepsilon_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)}, A_{t,i} = k\right] + o_p(1)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}}{lf(n)} \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(\varepsilon_{t,i}^2 > \delta^2 \frac{l\, nf(n)^2}{a\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}})} \middle| H_{t-1}^{(n)}, A_{t,i} = k\right] + o_p(1)$$

It is sufficient to show that

$$\lim_{n\to\infty} \max_{i\in[1:\,n]} \frac{1}{f(n)} \mathbb{E}\left[\varepsilon_{t,i}^2 \mathbb{I}_{(\varepsilon_{t,i}^2 > \delta^2 \frac{l\, nf(n)^2}{a\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}})} \middle| H_{t-1}^{(n)}, A_{t,i} = k\right] = 0. \tag{A.5.9}$$

By Condition A.2.1, we have that for all $n \geq 1$, $\max_{t\in[1:\,T],i\in[1:\,n]} \mathbb{E}[\phi(\varepsilon_{t,i}^2)|H_{t-1}^{(n)}, A_{t,i} = k] < M$.

Since we assume that $\lim_{x\to\infty} \frac{\phi(x)}{x} = \infty$, for all $m \geq 1$, there exists a b_m s.t. $\phi(x) \geq mMx$ for all $x \geq b_m$. So, for all n, t, i,

$$M \geq \mathbb{E}[\phi(\varepsilon_{t,i}^2)|H_{t-1}^{(n)}, A_{t,i} = k] \geq \mathbb{E}[\phi(\varepsilon_{t,i}^2)\mathbb{I}_{(\varepsilon_{t,i}^2 \geq b_m)}|H_{t-1}^{(n)}, A_{t,i} = k]$$

$$\geq mM\mathbb{E}[\varepsilon_{t,i}^2 \mathbb{I}_{(\varepsilon_{t,i}^2 \geq b_m)}|H_{t-1}^{(n)}, A_{t,i} = k]$$

Thus, $\max_{i\in[1:\,n]} \mathbb{E}[\varepsilon_{t,i}^2 \mathbb{I}_{(\varepsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = k] \leq \frac{1}{m}$; so

$$\lim_{m\to\infty} \max_{i\in[1:\,n]} \mathbb{E}[\varepsilon_{t,i}^2 \mathbb{I}_{(\varepsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = k] = 0.$$

Since by our conditional clipping assumption, $f(n) = c$ for some $0 < c \leq \frac{1}{2}$ thus $nf(n)^2 \to \infty$. So equation (A.5.9) holds. □

Corollary A.5.1 (Asymptotic Normality of the Batched OLS for Margin with Context Statistic)**.** *Assume the same conditions as Theorem 3.5.2. For any two arms $x, y \in [0:\, K-1]$ for all $t \in [1:\, T]$, we have the BOLS estimator for $\Delta_{t,x-y} \triangleq \beta_{t,x} - \beta_{t,y}$. We show that as* $n \to \infty$,

$$\begin{bmatrix} \left[\underline{\boldsymbol{C}}_{1,x}^{-1} + \underline{\boldsymbol{C}}_{1,y}^{-1}\right]^{1/2} (\hat{\Delta}_{1,x-y}^{\text{BOLS}} - \Delta_{1,x-y}) \\ \left[\underline{\boldsymbol{C}}_{2,x}^{-1} + \underline{\boldsymbol{C}}_{2,y}^{-1}\right]^{1/2} (\hat{\Delta}_{2,x-y}^{\text{BOLS}} - \Delta_{2,x-y}) \\ \vdots \\ \left[\underline{\boldsymbol{C}}_{T,x}^{-1} + \underline{\boldsymbol{C}}_{T,y}^{-1}\right]^{1/2} (\hat{\Delta}_{T,x-y}^{\text{BOLS}} - \Delta_{T,x-y}) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_{Td})$$

where

$$\hat{\Delta}_{t,x-y}^{\text{BOLS}} = \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{-1} \left(\underline{\mathbf{C}}_{t,y}^{-1} \sum_{i=1}^{n} A_{t,i} \mathbf{C}_{t,i} R_{t,i} - \underline{\mathbf{C}}_{t,x}^{-1} \sum_{i=1}^{n} (1 - A_{t,i}) \mathbf{C}_{t,i} R_{t,i} \right).$$

Proof: By Cramer-Wold device, it is sufficient to show that for any fixed vector $\mathbf{d} \in \mathbb{R}^{Td}$ s.t. $\|\mathbf{d}\|_2 = 1$, where $\mathbf{d} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_T]$ for $\mathbf{d}_t \in \mathbb{R}^d$, $\sum_{t=1}^{T} \mathbf{d}_t^\top \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{1/2} (\hat{\Delta}_{t,x-y}^{\text{BOLS}} -$

$\Delta_{t,x-y}) \xrightarrow{D} \mathcal{N}(0,\sigma^2)$, as $n \to \infty$.

$$\sum_{t=1}^{T} \mathbf{d}_t^\top \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{1/2} (\hat{\Delta}_{t,x-y}^{\text{BOLS}} - \Delta_{t,x-y})$$
$$= \sum_{t=1}^{T} \mathbf{d}_t^\top \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{-1/2} \left(\underline{\mathbf{C}}_{t,y}^{-1} \sum_{i=1}^{n} A_{t,i}\mathbf{C}_{t,i}\varepsilon_{t,i} - \underline{\mathbf{C}}_{t,x}^{-1} \sum_{i=1}^{n} (1-A_{t,i})\mathbf{C}_{t,i}\varepsilon_{t,i}\right)$$

By Lemma A.5.1, as $n \to \infty$, $\frac{1}{nP_{t,x}}\underline{\mathbf{Z}}_{t,x}^{-1}\underline{\mathbf{C}}_{t,x} \xrightarrow{P} \underline{\mathbf{I}}_d$ and $\frac{1}{nP_{t,y}}\underline{\mathbf{Z}}_{t,y}^{-1}\underline{\mathbf{C}}_{t,y} \xrightarrow{P} \underline{\mathbf{I}}_d$, so by Slutsky's Theorem it is sufficient to that as $n \to \infty$,

$$\sum_{t=1}^{T} \mathbf{d}_t^\top \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{-1/2} \left(\frac{1}{nP_{t,y}}\underline{\mathbf{Z}}_{t,y}^{-1} \sum_{i=1}^{n} A_{t,i}\mathbf{C}_{t,i}\varepsilon_{t,i} - \frac{1}{nP_{t,x}}\underline{\mathbf{Z}}_{t,x}^{-1} \sum_{i=1}^{n} (1-A_{t,i})\mathbf{C}_{t,i}\varepsilon_{t,i}\right) \xrightarrow{D} \mathcal{N}(0,\sigma^2)$$

We know that $\left[\frac{1}{P_{t,x}}\underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{P_{t,y}}\underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \left[\frac{1}{P_{t,x}}\underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{P_{t,y}}\underline{\mathbf{Z}}_{t,y}^{-1}\right]^{1/2} \xrightarrow{P} \underline{\mathbf{I}}_d$.

By Lemma A.5.1 and continuous mapping theorem, we have that $nP_{t,y}\underline{\mathbf{Z}}_{t,y}\underline{\mathbf{C}}_{t,y}^{-1} \xrightarrow{P} \underline{\mathbf{I}}_d$ and $nP_{t,x}\underline{\mathbf{Z}}_{t,x}\underline{\mathbf{C}}_{t,x}^{-1} \xrightarrow{P} \underline{\mathbf{I}}_d$. So by Slutsky's Theorem,

$$\left[\frac{1}{P_{t,x}}\underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{P_{t,y}}\underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \left[n\underline{\mathbf{C}}_{t,x}^{-1} + n\underline{\mathbf{C}}_{t,y}^{-1}\right]^{1/2} \xrightarrow{P} \underline{\mathbf{I}}_d$$

So, returning to our CLT, by Slutsky's Theorem, it is sufficient to show that as $n \to \infty$,

$$\sum_{t=1}^{T} \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}}\underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}}\underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \frac{1}{nP_{t,y}}\underline{\mathbf{Z}}_{t,y}^{-1} \sum_{i=1}^{n} A_{t,i}\mathbf{C}_{t,i}\varepsilon_{t,i}$$

$$-\sum_{t=1}^{T} \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}}\underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}}\underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \frac{1}{nP_{t,x}}\underline{\mathbf{Z}}_{t,x}^{-1} \sum_{i=1}^{n} (1-A_{t,i})\mathbf{C}_{t,i}\varepsilon_{t,i} \xrightarrow{D} \mathcal{N}(0,\sigma^2)$$

The above sum equals the following:

$$= \sum_{t=1}^{T} \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1} \right]^{-1/2} \frac{1}{\sqrt{nP_{t,x}}} \underline{\mathbf{Z}}_{t,x}^{-1/2} \left(\frac{1}{\sqrt{nP_{t,x}}} \underline{\mathbf{Z}}_{t,x}^{-1/2} \sum_{i=1}^{n} A_{t,i} \mathbf{C}_{t,i} \varepsilon_{t,i} \right)$$

$$- \sum_{t=1}^{T} \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{\sqrt{nP_{t,x}}} \underline{\mathbf{Z}}_{t,y}^{-1} \right]^{-1/2} \frac{1}{\sqrt{nP_{t,y}}} \underline{\mathbf{Z}}_{t,y}^{-1/2} \left(\frac{1}{\sqrt{nP_{t,y}}} \underline{\mathbf{Z}}_{t,y}^{-1/2} \sum_{i=1}^{n} (1 - A_{t,i}) \mathbf{C}_{t,i} \varepsilon_{t,i} \right)$$

Asymptotic normality holds by the same martingale CLT as we used in the proof of Theorem 3.5.2. The only difference is that we adjust our $\mathbf{b}_{t,k}$ vector from Theorem 3.5.2 to the following:

$$\mathbf{b}_{t,k} \triangleq \begin{cases} 0 & \text{if } k \notin \{x, y\} \\ \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1} \right]^{-1/2} \frac{1}{\sqrt{nP_{t,x}}} \underline{\mathbf{Z}}_{t,x}^{-1/2} & \text{if } k = x \\ \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1} \right]^{-1/2} \frac{1}{\sqrt{nP_{t,y}}} \underline{\mathbf{Z}}_{t,y}^{-1/2} & \text{if } k = y \end{cases}$$

The proof still goes through with this adjustment because for all $k \in [0 : K-1]$, (i) $\mathbf{b}_{t,k} \in H_{t-1}^{(n)}$, (ii) $\sum_{t=1}^{T} \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \mathbf{b}_{t,k} = \sum_{t=1}^{T} \mathbf{d}_t^\top \mathbf{d}_t = 1$. and (iii) $\frac{l \, nf(n)^2}{a \mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}} \to \infty$ still holds because $\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}$ is bounded above by one. $\square$

A.6 W-Decorrelated Estimator

To better understand why the W-decorrelated estimator has relatively low power, but is still able to guarantee asymptotic normality, we now investigate the form of the W-decorrelated estimator in the two-arm bandit setting.

A.6.1 Decorrelation Approach

We now consider the unbatched setting (i.e., batch size is one), as the W-decorrelated estimator was developed for this setting; however, these results easily translate to the batched setting. We now let n index the number of samples total (previously this was nT) and examine asymptotics as $n \to \infty$. We assume the following model:

$$\mathbf{R}_n = \underline{\mathbf{X}}_n^\top \beta + \varepsilon_n$$

where $\mathbf{R}_n, \varepsilon_n \in \mathbb{R}^n$ and $\underline{\mathbf{X}}_n \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$. The W-decorrelated OLS estimator is defined as follows:

$$\hat{\beta}^d = \hat{\beta}_{\text{OLS}} + \underline{\mathbf{W}}_n(\mathbf{R}_n - \underline{\mathbf{X}}_n\hat{\beta}_{\text{OLS}})$$

With this definition we have that,

$$\hat{\beta}^d - \beta = \hat{\beta}_{\text{OLS}} + \underline{\mathbf{W}}_n(\mathbf{R}_n - \underline{\mathbf{X}}_n\hat{\beta}_{\text{OLS}}) - \beta$$

$$= \hat{\beta}_{\text{OLS}} + \underline{\mathbf{W}}_n(\underline{\mathbf{X}}_n\beta + \varepsilon_n) - \underline{\mathbf{W}}_n\underline{\mathbf{X}}_n\hat{\beta}_{\text{OLS}} - \beta$$

$$= (\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n\underline{\mathbf{X}}_n)(\hat{\beta}_{\text{OLS}} - \beta) + \underline{\mathbf{W}}_n\varepsilon_n$$

Note that if $\mathbb{E}[\underline{\mathbf{W}}_n \varepsilon_n] = \mathbb{E}\big[\sum_{i=1}^n \mathbf{W}_i \varepsilon_i\big] = 0$ (where $\mathbf{W}_i$ is the i^{th} column of $\underline{\mathbf{W}}_n$), then $\mathbb{E}[(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n)(\hat{\beta}_{\text{OLS}} - \beta)]$ would be the bias of the estimator. We assume $\{\varepsilon_i\}$ is a martingale difference sequence w.r.t. filtration $\{\mathcal{G}_i\}_{i=1}^n$. Thus, if we constrain $\mathbf{W}_i$ to be $\mathcal{G}_{i-1}$ measurable,

$$\mathbb{E}[\underline{\mathbf{W}}_n \varepsilon_n] = \mathbb{E}\big[\sum_{i=1}^n \mathbf{W}_i \varepsilon_i\big] = \sum_{i=1}^n \mathbb{E}\Big[\mathbb{E}[\mathbf{W}_i \varepsilon_i | \mathcal{G}_{i-1}]\Big] = \sum_{i=1}^n \mathbb{E}\Big[\mathbf{W}_i \mathbb{E}[\varepsilon_i | \mathcal{G}_{i-1}]\Big] = 0$$

Trading off Bias and Variance While decreasing $\mathbb{E}[(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n)(\hat{\beta}_{\text{OLS}} - \beta)]$ will decrease the bias, making $\underline{\mathbf{W}}_n$ larger in norm will increase the variance. So the trade-off between bias and variance can be adjusted with different values of λ for the following optimization problem:

$$||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n||_F^2 + \lambda ||\underline{\mathbf{W}}_n||_F^2 = ||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n||_F^2 + \lambda \text{Tr}(\underline{\mathbf{W}}_n \underline{\mathbf{W}}_n^\top)$$

Optimizing for $\underline{\mathbf{W}}_n$ The authors propose to optimize for $\underline{\mathbf{W}}_n$ in a recursive fashion, so that the i^{th} column, $\mathbf{W}_i$, only depends on $\{\mathbf{X}_j\}_{j \le i} \cup \{\varepsilon_j\}_{j \le i-1}$ (so $\sum_{i=1}^n \mathbb{E}[\mathbf{W}_i \varepsilon_i] = 0$). We let $\mathbf{W}_0 = 0, \mathbf{X}_0 = 0$, and recursively define $\underline{\mathbf{W}}_n := [\underline{\mathbf{W}}_{n-1} \mathbf{W}_n]$ where

$$\mathbf{W}_n = \text{argmin}_{\mathbf{W} \in \mathbb{R}^p} ||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top||_F^2 + \lambda ||\mathbf{W}||_2^2$$

where $\underline{\mathbf{W}}_{n-1} = [\mathbf{W}_1; \mathbf{W}_2; ...; \mathbf{W}_{n-1}] \in \mathbb{R}^{p \times (n-1)}$ and $\underline{\mathbf{X}}_{n-1} = [\mathbf{X}_1; \mathbf{X}_2; ...; \mathbf{X}_{n-1}]^\top \in \mathbb{R}^{(n-1) \times p}$. Now, let us find the closed form solution for each step of this minimization:

$$\frac{d}{d\mathbf{W}} ||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top||_F^2 + \lambda ||\mathbf{W}||_2^2 = 2(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top)(-\mathbf{X}_n) + 2\lambda \mathbf{W}$$

Note that since the Hessian is positive definite, so we can find the minimizing $\mathbf{W}$ by setting the first derivative to 0:

$$\frac{d^2}{d\mathbf{W}d\mathbf{W}^\top}\|\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1}\underline{\mathbf{X}}_{n-1} - \mathbf{W}\mathbf{X}_n^\top\|_F^2 + \lambda\|\mathbf{W}\|_2^2 = 2\mathbf{X}_n\mathbf{X}_n^\top + 2\lambda\underline{\mathbf{I}}_p \succcurlyeq 0$$

$$0 = 2(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1}\underline{\mathbf{X}}_{n-1} - \mathbf{W}\mathbf{X}_n^\top)(-\mathbf{X}_n) + 2\lambda\mathbf{W}$$

$$(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1}\underline{\mathbf{X}}_{n-1} - \mathbf{W}\mathbf{X}_n^\top)\mathbf{X}_n = \lambda\mathbf{W}$$

$$(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1}\underline{\mathbf{X}}_{n-1})\mathbf{X}_n = \lambda\mathbf{W} + \mathbf{W}\mathbf{X}_n^\top\mathbf{X}_n = (\lambda + \|\mathbf{X}_n\|_2^2)\mathbf{W}$$

$$\mathbf{W}^* = (\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1}\underline{\mathbf{X}}_{n-1})\frac{\mathbf{X}_n}{\lambda + \|\mathbf{X}_n\|_2^2}$$

Proposition A.6.1 (W-decorrelated estimator and time discounting in the two-arm bandit setting)**.** *Suppose we have a 2-arm bandit. A_i is an indicator that equals 1 if arm 1 is chosen for the i^{th} sample, and 0 if arm 0 is chosen. We define $\boldsymbol{X}_i := [1 - A_i, A_i] \in \mathbb{R}^2$. We assume the following model of rewards:*

$$R_i = \boldsymbol{X}_i^\top\beta + \varepsilon_i = A_i\beta_1 + (1 - A_i)\beta_0 + \varepsilon_i$$

We further assume that $\{\varepsilon_i\}_{i=1}^n$ are a martingale difference sequence with respect to filtration $\{\mathcal{G}_i\}_{i=1}^n$. We also assume that $\boldsymbol{X}_i$ are non-anticipating with respect to filtration $\{\mathcal{G}_i\}_{i=1}^n$. Note the W-decorrelated estimator:

$$\hat{\beta}^d = \hat{\beta}_{\text{OLS}} + \underline{\boldsymbol{W}}_n(\boldsymbol{R}_n - \underline{\boldsymbol{X}}_n\hat{\beta}_{\text{OLS}})$$

We show that for $\underline{\boldsymbol{W}}_n = [\boldsymbol{W}_1; \boldsymbol{W}_2; ...; \boldsymbol{W}_n] \in \mathbb{R}^{p \times n}$ *and choice of constant* λ*,*

$$\boldsymbol{W}_i = \begin{bmatrix} (1 - \frac{1}{\lambda+1})^{\sum_{i=1}^n (1-A_i)} \frac{1}{\lambda+1} \\ (1 - \frac{1}{\lambda+1})^{\sum_{i=1}^n A_i} \frac{1}{\lambda+1} \end{bmatrix} \in \mathbb{R}^2$$

Moreover, we show that the W-decorrelated estimator for the mean of arm 1*,* β_1*, is as follows:*

$$\hat{\beta}_1^d = \left(1 - \sum_{i=1}^n A_t \frac{1}{\lambda+1}\left(1 - \frac{1}{\lambda+1}\right)^{N_{1,i}-1}\right)\hat{\beta}_1^{\text{OLS}} + \sum_{i=1}^n A_t R_t \cdot \frac{1}{\lambda+1}\left(1 - \frac{1}{\lambda+1}\right)^{N_{1,i}-1}$$

where $\hat{\beta}_1^{\text{OLS}} = \frac{\sum_{i=1}^n A_i R_i}{N_{1,n}}$ *for* $N_{1,n} = \sum_{i=1}^n A_i$*. Since*[26] *require that* $\lambda \geq 1$ *for their CLT results to hold, thus, the W-decorrelated estimators is down-weighting samples drawn later on in the study and up-weighting earlier samples.*

Proof: Recall the formula for $\mathbf{W}_i$,

$$\mathbf{W}_i = (\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{i-1}\underline{\mathbf{X}}_{i-1})\frac{\mathbf{X}_i}{\lambda + \|\mathbf{X}_i\|_2^2}$$

We let $\mathbf{W}_i = [W_{0,i}, W_{1,i}]^\top$. For notational simplicity, we let $r = \frac{1}{\lambda+1}$. We now solve for $W_{1,n}$:

$$W_{1,1} = (1 - 0) \cdot rA_1 = rA_1$$

$$W_{1,2} = (1 - W_{1,1} \cdot A_1) \cdot rA_2 = (1 - rA_1)rA_2$$

$$W_{1,3} = \left(1 - \sum_{i=1}^2 \mathbf{W}_{1,i} \cdot A_i\right) \cdot rA_3 = \left(1 - rA_1 - (1 - rA_1)rA_2\right) \cdot rA_3 = (1 - rA_1)(1 - rA_2) \cdot rA_3$$

$$W_{1,4} = \left(1-\sum_{i=1}^{3} \mathbf{W}_{1,i}\cdot A_i\right)\cdot rA_4 = \left(1-rA_1-(1-rA_1)rA_2-(1-rA_1)(1-rA_2)\cdot rA_3\right)\cdot rA_4$$

$$= (1-rA_1)\big(1-rA_2-(1-rA_2)rA_3)\big)\cdot rA_4 = (1-rA_1)(1-rA_2)(1-rA_3)\cdot rA_4$$

We have that for arbitrary n,

$$W_{1,n} = \left(1-\sum_{i=1}^{n-1} \mathbf{W}_{1,i}\cdot A_i\right)\cdot rA_n = rA_n\prod_{i=1}^{n-1}(1-rA_i) = rA_n(1-r)^{\sum_{i=1}^{n-1} A_i} = rA_n(1-r)^{N_{1,n-1}}$$

By symmetry, we have that

$$W_{0,n} = \left(1-\sum_{i=1}^{n-1} \mathbf{W}_{1,i}\cdot(1-A_i)\right)\cdot r(1-A_n) = r(1-A_n)(1-r)^{N_{0,n-1}}$$

Note the W-decorrelated estimator for β_1:

$$\hat{\beta}_1^d = \hat{\beta}_1^{\text{OLS}} + \sum_{i=1}^{n} A_i\left(R_i - \hat{\beta}_1^{\text{OLS}}\right)r(1-r)^{N_{1,i-1}}$$

$$= \left(1-\sum_{i=1}^{n} A_i r(1-r)^{N_{1,i-1}}\right)\hat{\beta}_1^{\text{OLS}} + \sum_{i=1}^{n} A_i R_i\cdot r(1-r)^{N_{1,i-1}} \quad \square$$

B

Adaptively Weighted M-Estimators

B.1 Simulations

B.1.1 Simulation Details

Simulation Environment

- Each dimension of X_t is sampled independently from Uniform$(0, 5)$.
- $\theta^*(\mathcal{P}) = [\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})] = [0.1, 0.1, 0.1, 0, 0, 0]$, where $\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P}) \in \mathbb{R}^3$.
  Below also include simulations where $[\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})] = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$.
- t-Distributed rewards: $R_t|X_t, A_t \sim t_5 + \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$, where t_5 is a t-distribution with 5 degrees of freedom.
- Bernoulli rewards: $R_t|X_t, A_t \sim$ Bernoulli$(expit(\nu_t))$ for $\nu_t = \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$ and $expit(x) = \frac{1}{1+\exp(-x)}$.
- Poisson rewards: $R_t|X_t, A_t \sim$ Poisson$(\exp(\nu_t))$ for $\nu_t = \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$.

Algorithm

- Thompson Sampling with $\mathcal{N}(0, I_d)$ priors on each arm.
- 0.05 clipping
- Pre-processing rewards before received by algorithm:
  - Bernoulli: $2R_t - 1$
  - Poisson: $0.6R_t$

Compute Time and Resources All simulations run within a few hours on a MacBook Pro.

B.1.2 Details on Constructing of Confidence Regions

For notational convenience, we define $Z_t = [\tilde{X}_t, A_t\tilde{X}_t]$.

Least Squares Estimators

- $\hat{\theta}_T = \left(\sum_{t=1}^T W_t Z_t Z_t^\top\right)^{-1} \sum_{t=1}^T W_t Z_t R_t$
  - For unweighted least squares, $W_t = 1$ and we call the estimator $\hat{\theta}_T^{\text{OLS}}$.
  - For adaptively weighted least squares, $W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$; this is equivalent to using square-root importance weights with a uniform stabilizing policy. We call the estimator $\hat{\theta}_T^{\text{AW-LS}}$.

- We assume homoskedastic errors and estimate the noise variance σ^2 as follows:

$$\hat{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^T (R_t - Z_t^\top \hat{\theta}_T)^2.$$

- We use a Hotelling t-squared test statistic to construct confidence regions for $\theta^*(\mathcal{P})$:

$$C_T(\alpha) = \left\{\theta \in \mathbb{R}^d : \left[\hat{\Sigma}_T^{-1/2}\left(\frac{1}{T}\sum_{t=1}^T W_t Z_t Z_t^\top\right)\sqrt{T}(\hat{\theta}_T - \theta)\right]^{\otimes 2} \leq \frac{d(T-1)}{T-d} F_{d,T-d}(1-\alpha)\right\}. \quad \text{(B.1.1)}$$

- For the unweighted least-squares estimator we use the following variance estimator: $\hat{\Sigma}_T = \hat{\sigma}_T^2 \frac{1}{T} \sum_{t=1}^{T} Z_t Z_t^\top$.
- For the AW-Least Squares estimator we use the following variance estimator: $\hat{\Sigma}_T = \hat{\sigma}_T^2 \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}^{A_t} \frac{1}{1-\pi_t(A_t, X_t, \mathcal{H}_{t-1})}^{1-A_t} Z_t Z_t^\top$.

- To construct (non-projected) confidence regions for $\theta_1^*(\mathcal{P}) \in \mathbb{R}^{d_1}$ we treat the unweighted least squares / AW-LS estimators, $\hat{\theta}_{T,1}$, as

$$\mathcal{N}\left(\theta_1^*(\mathcal{P}), \frac{1}{T}\left(\frac{1}{T}\sum_{t=1}^{T} W_t Z_t Z_t^\top\right)^{-1} \hat{\Sigma}_T \left(\frac{1}{T}\sum_{t=1}^{T} W_t Z_t Z_t^\top\right)^{-1}\right).$$

We use a Hotelling t-squared test statistic to construct confidence regions for $\theta_1^*(\mathcal{P})$:

$$C_T(\alpha) = \left\{\theta_1 \in \mathbb{R}^{d_1} : \left[V_{1,T}^{-1/2}\sqrt{T}(\hat{\theta}_{T,1} - \theta_1)\right]^{\otimes 2} \leq \frac{d_1(T-1)}{T-d_1} F_{d_1, T-d_1}(1-\alpha)\right\},$$

where $V_{1,T}$ is the lower right $d_1 \times d_1$ block of matrix

$$\left(\frac{1}{T}\sum_{t=1}^{T} W_t Z_t Z_t^\top\right)^{-1} \hat{\Sigma}_T \left(\frac{1}{T}\sum_{t=1}^{T} W_t Z_t Z_t^\top\right)^{-1}.$$

Recall that for the unweighted least squares estimator $W_t = 1$ and for AW-LS $W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$.

- For the AW-least squares estimator, we also construct projected confidence regions for $\theta_1^*(\mathcal{P})$ using the confidence region defined in equation (B.1.1). See Section B.1.2 below for more details on constructing projected confidence regions.

MLE Estimators

Distribution	ν	$b(\nu)$	$b'(\nu)$	$b''(\nu)$	$b'''(\nu)$
$\mathcal{N}(\mu, 1)$	μ	$\frac{1}{2}\nu^2$	$\nu = \mu$	1	0
Poisson(λ)	$\log \lambda$	$\exp(\nu)$	$\exp(\nu) = \lambda$	$\exp(\nu) = \lambda$	$\exp(\nu) = \lambda$
Bernoulli(p)	$\log\left(\frac{p}{1-p}\right)$	$\log(1 + e^\nu)$	$\frac{e^\nu}{1+e^\nu} = p$	$\frac{e^\nu}{(1+e^\nu)^2} = p(1-p)$	$p(1-p)(1-2p)$

- $\hat{\theta}_T$ is the root of the score function:

$$0 = \sum_{t=1}^{T} W_t \left(R_t - b'(\hat{\theta}_T^\top Z_t)\right) Z_t.$$

  We use Newton Raphson optimization to solve for $\hat{\theta}_T$.

  - For unweighted MLE, $W_t = 1$.
  - For AW-MLE, $W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$; this is equivalent to using square-root importance weights with a uniform stabilizing policy.

- Second derivative of score function: $-\sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$.

- We use a Hotelling t-squared test statistic to construct confidence regions for $\theta^*(\mathcal{P})$:

$$C_T(\alpha) = \left\{\theta \in \mathbb{R}^d : \left[\hat{\Sigma}_T^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top\right)\sqrt{T}(\hat{\theta}_T - \theta)\right]^{\otimes 2} \leq \frac{d(T-1)}{T-d} F_{d,T-d}(1-\alpha)\right\}. \quad \text{(B.1.2)}$$

  - For the MLE variance estimator, we use $\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$.

– For the AW-MLE variance estimator, we use

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}^{A_t} \frac{1}{1-\pi_t(A_t, X_t, \mathcal{H}_{t-1})}^{1-A_t} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top.$$

- To construct (non-projected) confidence regions for $\theta_1^*(\mathcal{P}) \in \mathbb{R}^{d_1}$ we treat the MLE / AW-MLE estimators, $\hat{\theta}_{T,1}$, as

$$\mathcal{N}\left(\theta_1^*(\mathcal{P}), \frac{1}{T}\left(\frac{1}{T}\sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top\right) \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top\right)\right).$$

We use a Hotelling t-squared test statistic to construct confidence regions for $\theta_1^*(\mathcal{P})$:

$$C_T(\alpha) = \left\{\theta_1 \in \mathbb{R}^{d_1} : \left[V_{1,T}^{-1/2}\sqrt{T}(\hat{\theta}_{T,1} - \theta_1)\right]^{\otimes 2} \leq \frac{d_1(T-1)}{T-d_1} F_{d_1, T-d_1}(1-\alpha)\right\},$$

where $V_{1,T}$ is the lower right $d_1 \times d_1$ block of matrix

$$\left(\frac{1}{T}\sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top\right) \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top\right).$$

- For the AW-MLE estimator, we also construct projected confidence regions for $\theta_1^*(\mathcal{P})$ using the confidence region defined in equation (B.1.2). See Section B.1.2 below for more details on constructing projected confidence regions.

W-Decorrelated

The following is based on Algorithm 1 of Deshpande et al. [26].

- The W-decorrelated estimator for $\theta^*(\mathcal{P})$ is constructed as follows with adaptive

weights for $W_t \in \mathbb{R}^d$:

$$\hat{\theta}_T^{\text{WD}} = \hat{\theta}_T^{\text{OLS}} + \sum_{t=1}^{T} W_t (R_t - \tilde{X}_t^\top \hat{\theta}_T^{\text{OLS}}).$$

- The weights are set as follows:

  $W_1 = 0 \in \mathbb{R}^d$ and $W_t = (I_d - \sum_{s=1}^{t} \sum_{u=1}^{t} W_s Z_u^\top) Z_t \frac{1}{\lambda_T + \|Z_t\|_2^2}$ for $t > 1$.

- We choose $\lambda_T = \text{mineig}_{0.01}(Z_t Z_t^\top)/\log T$ and $\text{mineig}_\alpha(Z_t Z_t^\top)$ represents the α quantile of the minimum eigenvalue of $Z_t Z_t^\top$. This is similar to the procedure used in the simulations of Deshpande et al. [26] and is guided by Proposition 5 in their paper.

- We assume homoskedastic errors and estimate the noise variance σ^2 as follows:

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_{t=1}^{T} (R_t - Z_t^\top \hat{\theta}_T^{\text{OLS}})^2.$$

- To construct confidence ellipsoids for $\theta^*(\mathcal{P})$ are constructed using a Hotelling t-squared statistic:

$$C_T(\alpha) = \left\{ \theta \in \mathbb{R}^d : (\hat{\theta}_T^{\text{WD}} - \theta)^\top V_T^{-1} (\hat{\theta}_T^{\text{WD}} - \theta) \leq \frac{d(T-1)}{T-d} F_{d,T-d}(1-\alpha) \right\}$$

  where $V_T = \hat{\sigma}_T^2 \sum_{t=1}^{T} W_t W_t^\top$.

- To construct confidence ellipsoids for $\theta_1^*(\mathcal{P}) \in \mathbb{R}^{d_1}$ with the following confidence

ellipsoid where $V_{T,1}$ is the lower right $d_1 \times d_1$ block of matrix V_T:

$$C_T(\alpha) = \left\{ \theta_1 \in \mathbb{R}^{d_1} : (\hat{\theta}^{\text{WD}}_{T,1} - \theta_1)^\top V_{T,1}^{-1} (\hat{\theta}^{\text{WD}}_{T,1} - \theta_1) \leq \frac{d_1(T-1)}{T-d_1} F_{d_1, T-d_1}(1-\alpha) \right\}.$$

Self-Normalized Martingale Bound

We construct $1-\alpha$ confidence region using the following equation taken from Theorem 2 of[2]:

$$C_T(\alpha) = \Bigg\{ \theta \in \Theta : (\hat{\theta}_T - \theta)^\top V_T (\hat{\theta}_T - \theta) \\ \leq \sigma \sqrt{2 \log \left(\frac{\det(V_T)^{1/2} \det(\lambda I_d)^{-1/2}}{\alpha} \right)} + \lambda^{1/2} S \Bigg\}.$$

- $\hat{\theta}_T = \left(\lambda I_d + \sum_{t=1}^{T} Z_t Z_t^\top \right)^{-1} \sum_{t=1}^{T} Z_t R_t$.
- $V_T = I_d \lambda + \sum_{t=1}^{T} Z_t Z_t^\top$.
- $\lambda = 1$ (ridge regression regularization parameter).
- $\sigma = 1$ (assumes rewards are σ-subgaussian).
- $S = 6$, where it is assumed that $\|\theta^*(\mathcal{P})\| \leq S$ (recall that in our simulations $\theta^*(\mathcal{P}) \in \mathbb{R}^6$).
- $\Theta = \{\theta \in \mathbb{R}^6 : \|\theta\|_2 \leq 6\}$.
- For constructing confidence regions for $\theta^*(\mathcal{P})$, we use projected confidence regions.

Construction of Projected Confidence Regions

We are interested in getting the confidence ellipsoid of the projection of a d-dimensional ellipsoid onto p-dimensional space, for $p < d$.

- Defining the original d-dimensional ellipsoid, for $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{B} \in \mathbb{R}^{d \times d}$:

$$\mathbf{x}^\top \mathbf{B}\mathbf{x} = 1$$

- Partitioning the matrix $\mathbf{B}$ and vector $\mathbf{x}$:

  For $y \in \mathbb{R}^{d-p}$ and $z \in \mathbb{R}^p$.

$$\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}$$

  For $\mathbf{C} \in \mathbb{R}^{d-p \times d-p}$, $\mathbf{E} \in \mathbb{R}^{p \times p}$, and $\mathbf{D} \in \mathbb{R}^{d-p \times p}$.

$$\mathbf{B} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{E} \end{bmatrix}$$

- Gradient of $\mathbf{x}^\top \mathbf{B}\mathbf{x}$ with respect to $\mathbf{x}$:

$$(\mathbf{B} + \mathbf{B}^\top)\mathbf{x} = 2\mathbf{B}\mathbf{x} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}.$$

  Since we are projecting onto the p-dimensional space, our projection is such that the

gradient of $\mathbf{x}^\top \mathbf{B}\mathbf{x}$ with respect to $\mathbf{y}$ is zero, which means

$$\mathbf{C}\mathbf{y} + \mathbf{D}\mathbf{z} = 0.$$

This means in the projection that $\mathbf{y} = -\mathbf{C}^{-1}\mathbf{D}\mathbf{z}$.

- Returning to our definition of the ellipsoid, plugging in $\mathbf{z}$, we have that

$$1 = \mathbf{x}^\top \mathbf{B}\mathbf{x} = \begin{bmatrix} \mathbf{y}^\top & \mathbf{z}^\top \end{bmatrix} \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} = \mathbf{y}^\top \mathbf{C}\mathbf{y} + 2\mathbf{z}^\top \mathbf{D}^\top \mathbf{y} + \mathbf{z}^\top \mathbf{E}\mathbf{z}$$

$$= (\mathbf{C}^{-1}\mathbf{D}\mathbf{z})^\top \mathbf{C}(\mathbf{C}^{-1}\mathbf{D}\mathbf{z}) - 2\mathbf{z}^\top \mathbf{D}^\top (\mathbf{C}^{-1}\mathbf{D}\mathbf{z}) + \mathbf{z}^\top \mathbf{E}\mathbf{z}$$

$$= \mathbf{z}^\top \mathbf{D}^\top \mathbf{C}^{-1}\mathbf{D}\mathbf{z} - 2\mathbf{z}^\top \mathbf{D}^\top \mathbf{C}^{-1}\mathbf{D}\mathbf{z} + \mathbf{z}^\top \mathbf{E}\mathbf{z}$$

$$= \mathbf{z}^\top (\mathbf{E} - \mathbf{D}^\top \mathbf{C}^{-1}\mathbf{D})\mathbf{z}.$$

Thus the equation for the final projected ellipsoid is

$$\mathbf{z}^\top (\mathbf{E} - \mathbf{D}^\top \mathbf{C}^{-1}\mathbf{D})\mathbf{z} = 1.$$

B.1.3 Additional Simulation Results

In addition to the continuous reward and a binary reward settings, here we also consider a discrete count reward setting. In this discrete reward setting, the reward R_t is generated from a Poisson distribution where $\mathbb{E}_{\mathcal{P}}[R_t | X_t, A_t] = \exp(\tilde{X}_t^\top \theta_0^*(\mathcal{P}) - A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P}))$. All other data generation methods are equivalent to those used for the other simulation set-

tings. Additionally we will consider the setting in which $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$ for the continuous reward, binary reward, and discrete count settings.

To analyze the data, in the discrete count reward setting, we assume a correctly specified model for the expected reward. We use both unweighted and adaptively weighted maximum likelihood estimators (MLEs), i.e., M-estimators with $m_\theta(R_t, X_t, A_t)$ set to the negative log-likelihood of R_t given X_t, A_t. We solve for these estimators using Newton–Raphson optimization and do not put explicit bounds on the parameter space Θ.

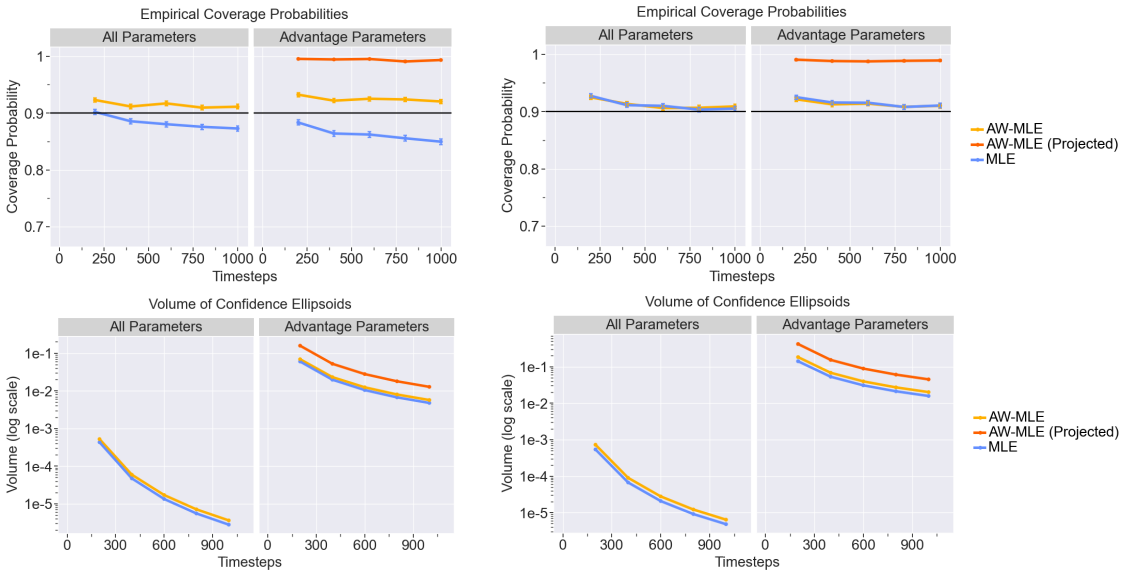

Figure B.1: Poisson Rewards: Empirical coverage probabilities for 90% confidence ellipsoids for parameters $\theta^*(\mathcal{P})$ and $\theta_1^*(\mathcal{P})$ (top row). We also plot the volumes of these 90% confidence ellipsoids for $\theta^*(\mathcal{P})$ and parameters $\theta_1^*(\mathcal{P})$ (bottom row). We set $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0, 0, 0]$ (left) and to $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$ (right).

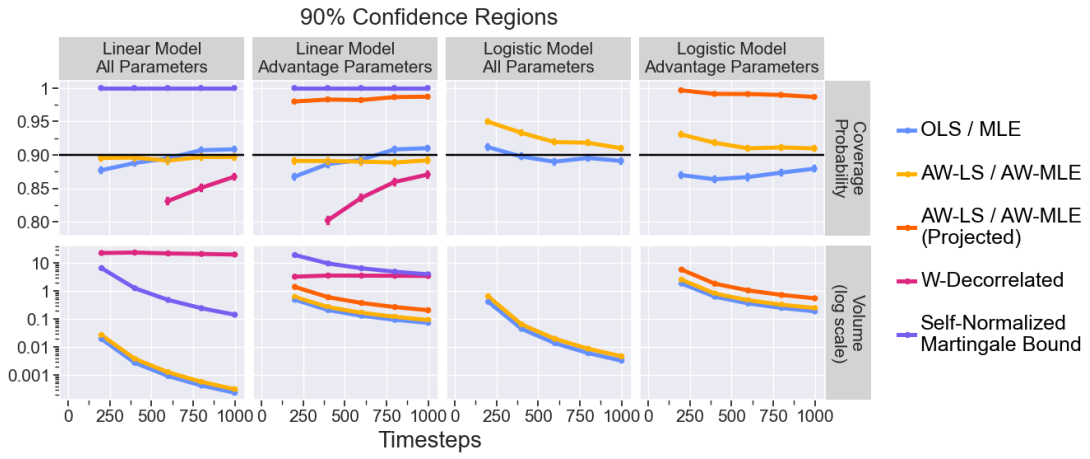

Figure B.2: Empirical coverage probabilities (upper row) and volume (lower row) of 90% confidence ellipsoids. In these simulations, $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$. The left two columns are for the linear reward model setting (t-distributed rewards) and the right two columns are for the logistic regression model setting (Bernoulli rewards). We consider confidence ellipsoids for all parameters $\theta^*(\mathcal{P})$ and for advantage parameters $\theta_1^*(\mathcal{P})$ for both settings.

In Figure B.3, we plot the mean squared errors of all estimators for all three simulation settings (same simulation hyperparameters as described previously for the respective simulation settings).

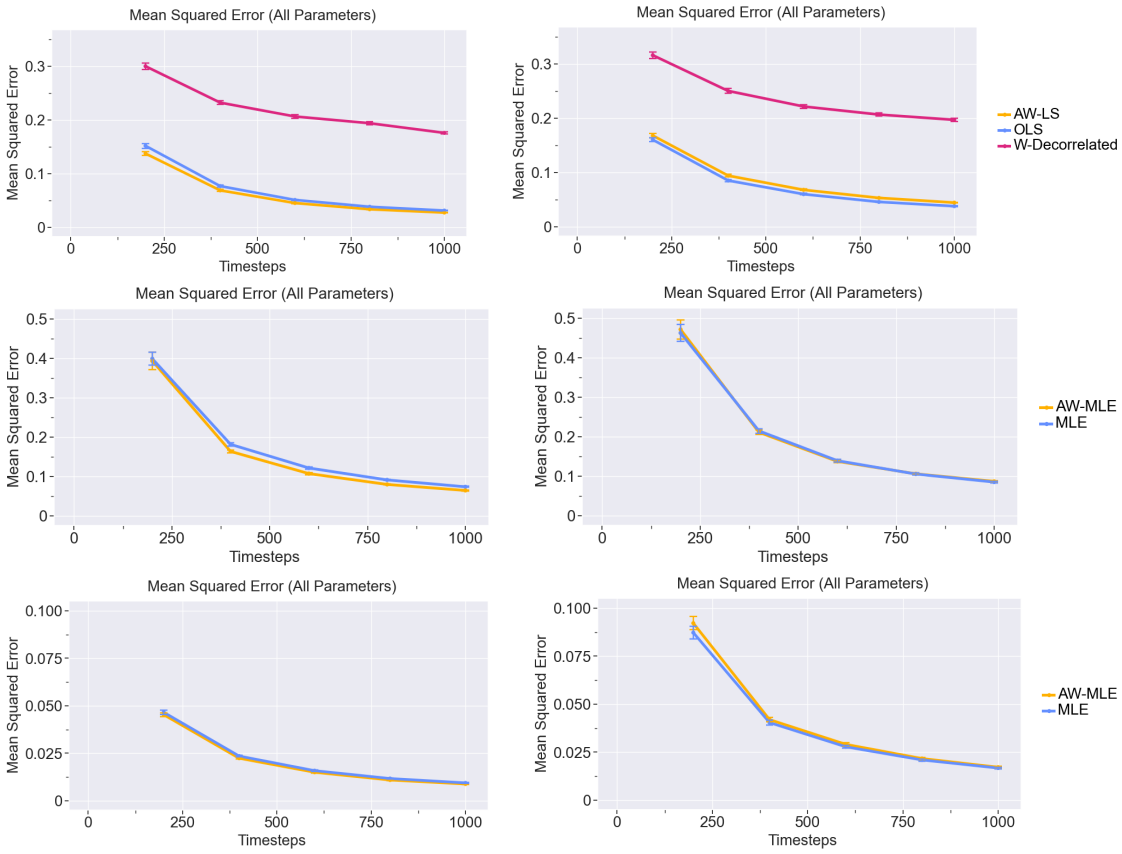

Figure B.3: Mean squared error estimators of $\theta^*(\mathcal{P})$ for linear model (top), logistic regression model (middle), and generalized linear model for Poisson rewards (bottom). We consider simulations with $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0, 0, 0]$ (left) and simulations with $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$ (right).

B.2 Asymptotic Results

Throughout, $\|\cdot\|$ refers to the L_2 norm.

B.2.1 Definitions

Here we define convergence in probability and distribution that is uniform over the true parameter. We follow the definitions are based on those in Kasy [55] and Van Der Vaart & Wellner [100], Chapter 1.12.

Definition B.2.1 (Uniform Convergence in Probability). *Let $\{Z_T(\mathcal{P})\}_{T\geq 1}$ be a sequence of random variables whose distributions are defined by some $\mathcal{P} \in \boldsymbol{P}$ and some nuisance component η. We say that $Z_T(\mathcal{P}) \xrightarrow{p} c$ uniformly over $\mathcal{P} \in \boldsymbol{P}$ as $T \to \infty$ if for any $\varepsilon > 0$,*

$$\sup_{\mathcal{P}\in\boldsymbol{P}} \mathbb{P}_{\mathcal{P},\eta} \left(\|Z_T(\mathcal{P}) - c\| > \varepsilon \right) \to 0. \tag{B.2.1}$$

For simplicity of notation, throughout we denote $Z_T(\mathcal{P}) - c = o_{\mathcal{P}\in\boldsymbol{P}}(1)$ to mean $Z_T(\mathcal{P}) \xrightarrow{p} c$ uniformly over $\mathcal{P} \in \boldsymbol{P}$ as $T \to \infty$.

Definition B.2.2 (Uniformly Stochastically Bounded). *Let $\{Z_T(\mathcal{P})\}_{T\geq 1}$ be a sequence of random variables whose distributions are defined by some $\mathcal{P} \in \boldsymbol{P}$ and some nuisance component η. We say that $Z_T(\mathcal{P})$ is uniformly stochastically bounded over $\mathcal{P} \in \boldsymbol{P}$ as $T \to \infty$ if for any $\varepsilon > 0$ there exists some $k < \infty$ such that*

$$\limsup_{T\to\infty} \sup_{\mathcal{P}\in\boldsymbol{P}} \mathbb{P}_{\mathcal{P},\eta} \left(\|Z_T(\mathcal{P})\| > k \right) < \varepsilon.$$

Similarly we denote $Z_T(P) = O_{\mathcal{P}\in\boldsymbol{P}}(1)$ *to mean* $Z_T(\mathcal{P})$ *is stochastically bounded uniformly over* $\mathcal{P} \in \boldsymbol{P}$ *as* $T \to \infty$.

Definition B.2.3 (Uniform Convergence in Distribution)**.** *Let* $Z(\mathcal{P}) \in \mathbb{R}^{d_Z}$ *and let* $\{Z_T(\mathcal{P})\}_{T\geq 1} \in \mathbb{R}^{d_Z}$ *be a sequence of random variables whose distributions are defined by some* $\mathcal{P} \in \boldsymbol{P}$ *and some nuisance component* η*. We say that* $Z_T(\mathcal{P}) \xrightarrow{D} Z(\mathcal{P})$ *uniformly over* $\mathcal{P} \in \boldsymbol{P}$ *as* $T \to \infty$ *if*

$$\sup_{\mathcal{P}\in\boldsymbol{P}} \sup_{f\in BL_1} \left| \mathbb{E}_{\mathcal{P},\eta}\left[f(Z_T(\mathcal{P}))\right] - \mathbb{E}_{\mathcal{P},\eta}\left[f(Z(\mathcal{P}))\right] \right| \to 0, \tag{B.2.2}$$

where BL_1 *is the set of functions* $f : \mathbb{R}^{d_z} \to \mathbb{R}$ *with* $\|f(z)\|_\infty \leq 1$ *and* $|f(z) - f(z')| \leq \|z - z'\|$ *for all* $z, z' \in \mathbb{R}^{d_Z}$.

As discussed in Kasy [55], Equation (B.2.1) holds if and only if for any $\varepsilon > 0$ and any sequence $\{\mathcal{P}_T\}_{T\geq 1}$ such that $\mathcal{P}_T \in \mathbf{P}$ for all $T \geq 1$, $\mathbb{P}_{\mathcal{P}_T,\eta}\left(\|Z_T(\mathcal{P}_T) - c\| > \varepsilon\right) \to 0$.

Similarly, Equation (B.2.2) holds if and only if for any sequence $\{\mathcal{P}_T\}_{T\geq 1}$ such that $\mathcal{P}_T \in \mathbf{P}$ for all $T \geq 1$, $\sup_{f\in BL_1} \left| \mathbb{E}_{\mathcal{P}_T,\eta}\left[f(Z_T(\mathcal{P}_T))\right] - \mathbb{E}_{\mathcal{P}_T,\eta}\left[f(Z(\mathcal{P}_T))\right] \right| \to 0$.

B.2.2 Consistency

We prove the first part of Theorem 4.3.1, i.e., that $\hat{\theta}_T \xrightarrow{p} \theta^*(\mathcal{P})$ uniformly over $\mathcal{P} \in \mathbf{P}$. We abbreviate $m_\theta(Y_t, X_t, A_t)$ with $m_{\theta,t}$. By definition of $\hat{\theta}_T$,

$$\sum_{t=1}^{T} W_t m_{\hat{\theta}_T,t} = \sup_{\theta\in\Theta} \sum_{t=1}^{T} W_t m_{\theta,t} \geq \sum_{t=1}^{T} W_t m_{\theta^*(\mathcal{P}),t}.$$

Note that $\|\hat{\theta}_T - \theta^*(\mathcal{P})\| > \varepsilon > 0$ implies that

$$\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \sum_{t=1}^{T} W_t m_{\theta,t} = \sup_{\theta\in\Theta} \sum_{t=1}^{T} W_t m_{\theta,t}.$$

Thus, the above two results imply the following inequality:

$$\sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\|\hat{\theta}_T - \theta^*(\mathcal{P})\| > \varepsilon\right)$$

$$\leq \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \sum_{t=1}^{T} W_t m_{\theta,t} \geq \sum_{t=1}^{T} W_t m_{\theta^*(\mathcal{P}),t}\right)$$

$$= \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t}\right\} - \frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta^*(\mathcal{P}),t} \geq 0\right)$$

$$= \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\Bigg[\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}] + \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\}$$
$$- \frac{1}{T}\sum_{t=1}^{T} \left\{W_t m_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}] + \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}]\right\} \geq 0\Bigg].$$

By triangle inequality,

$$\leq \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\Bigg(\underbrace{\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left(W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}] \right) \right\}}_{(a)}$$

$$+ \underbrace{\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}\left[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}) \middle| \mathcal{H}_{t-1} \right] \right\}}_{(b)}$$

$$- \underbrace{\frac{1}{T}\sum_{t=1}^{T} \left\{ W_t m_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}] \right\}}_{(c)} \geq 0 \Bigg) \to 0. \quad \text{(B.2.3)}$$

We now show that the limit in Equation (B.2.3) above holds.

- Regarding term (c), by moment bounds of Condition 4.3.5 and Lemma B.2.1, $\frac{1}{T}\sum_{t=1}^{T} \left\{ W_t m_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}] \right\} = o_{\mathcal{P}\in\mathbf{P}}(1)$.

- Regarding term (a), by Lemma B.2.2, $\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left(W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}] \right) \right\} = o_{\mathcal{P}\in\mathbf{P}}(1)$.

Thus it is sufficient to show that term (b) is such that for some $\delta' > 0$,

$$\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t})|\mathcal{H}_{t-1}] \right\} \leq -\delta' \text{ w.p. } 1. \quad \text{(B.2.4)}$$

By law of iterated expectations,

$$\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t})|\mathcal{H}_{t-1}] \right\}$$

$$= \sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t(a, X_t, \mathcal{H}_{t-1}) \right.\right.$$
$$\left.\left. \mathbb{E}_{\mathcal{P}}[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t})|\mathcal{H}_{t-1}, X_t, A_t = a]da \,\middle|\, \mathcal{H}_{t-1}\right]\right\}.$$

Since $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$, we have that $\mathbb{E}_{\mathcal{P}}[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t})|\mathcal{H}_{t-1}, X_t, A_t = a] = W_t\mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}, X_t, A_t = a]$. By Condition 4.3.1, we have that $W_t\mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}, X_t, A_t = a] = W_t\mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|X_t, A_t = a]$. Thus we have,

$$= \sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t(a, X_t, \mathcal{H}_{t-1})\, W_t \right.\right.$$
$$\left.\left. \mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|X_t, A_t = a]da \,\middle|\, \mathcal{H}_{t-1}\right]\right\}.$$

Since for all $\theta \in \Theta$, $\mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|X_t, A_t] \leq 0$ with probability 1 by Condition 4.3.7 and since $0 < \frac{W_t}{\sqrt{\rho_{\max}}} \leq 1$ with probability 1 by Condition 4.3.9,

$$\leq \sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t(a, X_t, \mathcal{H}_{t-1})\, W_t^2 \right.\right.$$
$$\left.\left. \mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|X_t, A_t = a]da \,\middle|\, \mathcal{H}_{t-1}\right]\right\}.$$

Since $W_t^2 = \frac{\pi_t^{\text{sta}}(A_t,X_t)}{\pi_t(A_t,X_t,\mathcal{H}_{t-1})}$,

$$= \sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\varepsilon} \left\{ \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \right.\right.$$
$$\left.\left. \mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|X_t, A_t = a]da \,\middle|\, \mathcal{H}_{t-1}\right]\right\}.$$

By Condition 4.3.1 and since π_t^{sta} is pre-specified, we can drop the conditioning on $\mathcal{H}_{t-1}$, i.e.,

$$= \sup_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\| > \varepsilon} \left\{ \frac{1}{T\sqrt{\rho_{\max}}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t} | X_t, A_t = a] da \right] \right\}.$$

By law of iterated expectations,

$$= \sup_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\| > \varepsilon} \left\{ \frac{1}{T\sqrt{\rho_{\max}}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t} \right] \right\} \leq -\frac{1}{\sqrt{\rho_{\max}}} \delta.$$

The last inequality above holds for some $\delta > 0$ for all sufficiently large T by Condition 4.3.8. Thus Equation (B.2.4) holds for $\delta' = \frac{1}{\sqrt{\rho_{\max}}} \delta$.

B.2.3 Asymptotic Normality

We prove the second part of Theorem 4.3.1, i.e., that

$$\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\hat{\theta}_T) \sqrt{T} (\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{B.2.5}$$

Main Argument

The three results we show to ensure Equation (B.2.5) holds are as follows:

$$\Sigma_T(\mathcal{P})^{-1/2} \sqrt{T} \dot{M}_T(\theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{B.2.6}$$

For $\dddot{\varepsilon}_{\ddot{m}} > 0$ as defined in Condition 4.3.6,

$$\sup_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\| \leq \varepsilon_{\ddot{m}}} \left\| \dddot{M}_T(\theta) \right\|_1 = O_{\mathcal{P} \in \mathbf{P}}(1). \tag{B.2.7}$$

For matrix H positive definite,

$$-\ddot{M}_T(\theta^*(\mathcal{P})) \succeq H + o_{\mathcal{P} \in \mathbf{P}}(1). \tag{B.2.8}$$

For a reminder on the notation of $o_{\mathcal{P} \in \mathbf{P}}(1)$ and $O_{\mathcal{P} \in \mathbf{P}}(1)$ see definitions B.2.1 and B.2.2. For now, we assume that Equations (B.2.6), (B.2.7), and (B.2.8) hold; we will show they hold in Sections B.2.3, B.2.3, and B.2.3 respectively. Our argument is based on Van der Vaart [99], Theorem of 5.41.

By differentiability Condition 4.3.2, since $\hat{\theta}_T$ is the maximizer of criterion $M_T(\theta)$,

$$0 = \dot{M}_T(\hat{\theta}_T).$$

By differentiability Condition 4.3.2 again and Taylor's theorem we have that for some random $\tilde{\theta}_T$ on the line segment between $\theta^*(\mathcal{P})$ and $\hat{\theta}_T$,

$$\begin{aligned} 0 = \dot{M}_T(\hat{\theta}_T) = \dot{M}_T(\theta^*(\mathcal{P})) + \ddot{M}_T(\theta^*(\mathcal{P}))(\hat{\theta}_T - \theta^*(\mathcal{P})) \\ + \frac{1}{2}(\hat{\theta}_T - \theta^*(\mathcal{P}))^\top \dddot{M}_T(\tilde{\theta}_T)(\hat{\theta}_T - \theta^*(\mathcal{P})). \end{aligned}$$

By rearranging terms and multiplying by $\sqrt{T}$,

$$-\sqrt{T}\dot{M}_T(\theta^*(\mathcal{P})) = \ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) + \frac{1}{2}(\hat{\theta}_T - \theta^*(\mathcal{P}))^\top \dddot{M}_T(\tilde{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P}))$$

$$= \left[\ddot{M}_T(\theta^*(\mathcal{P})) + \frac{1}{2}(\hat{\theta}_T - \theta^*(\mathcal{P}))^\top \dddot{M}_T(\tilde{\theta}_T)\right]\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})).$$

Note that by the above equation and Equation (B.2.6), we have that

$$\Sigma_T(\mathcal{P})^{-1/2}\left[\ddot{M}_T(\theta^*(\mathcal{P})) + \frac{1}{2}(\hat{\theta}_T - \theta^*(\mathcal{P}))^\top \dddot{M}_T(\tilde{\theta}_T)\right]\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \quad \text{(B.2.9)}$$

By Equation (B.2.8), the probability that $\ddot{M}_T(\theta^*(\mathcal{P}))$ is invertible goes to 1 uniformly over $\mathcal{P} \in \mathbf{P}$. Thus by Equation (B.2.9), we have that

$$\Sigma_T(\mathcal{P})^{-1/2}\left[I_d + \frac{1}{2}(\hat{\theta}_T - \theta^*(\mathcal{P}))^\top \dddot{M}_T(\tilde{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\right]\ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P}))$$

$$= \left[I_d + \frac{1}{2}\Sigma_T(\mathcal{P})^{-1/2}(\hat{\theta}_T - \theta^*(\mathcal{P}))^\top \dddot{M}_T(\tilde{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\Sigma_T(\mathcal{P})^{1/2}\right]$$
$$\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \quad \text{(B.2.10)}$$

We now show that $\frac{1}{2}\Sigma_T(\mathcal{P})^{-1/2}(\hat{\theta}_T - \theta^*(\mathcal{P}))^\top \dddot{M}_T(\tilde{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\Sigma_T(\mathcal{P})^{1/2} =$

$o_{\mathcal{P}\in\mathbf{P}}(1)$. It is sufficient to show that

$$\|\Sigma_T(\mathcal{P})^{-1/2}\|\|\hat{\theta}_T - \theta^*(\mathcal{P})\|\|\dddot{M}_T(\tilde{\theta}_T)\|_1\|\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\|\|\Sigma_T(\mathcal{P})^{1/2}\| = o_{\mathcal{P}\in\mathbf{P}}(1).$$

- By Condition 4.3.5, the minimum eigenvalue of $\Sigma_T(\mathcal{P})$ is bounded uniformly above some constant greater than zero, so $\sup_{\mathcal{P}\in\mathbf{P}} \|\Sigma_T(\mathcal{P})^{-1/2}\| = O(1)$.
- By uniform consistency of $\hat{\theta}_T$, $\|\hat{\theta}_T - \theta^*(\mathcal{P})\| = o_{\mathcal{P}\in\mathbf{P}}(1)$.
- By uniform consistency of $\hat{\theta}_T$, $1_{\|\tilde{\theta}_T - \theta^*(\mathcal{P})\| \leq \varepsilon} \dddot{M}_T = o_{\mathcal{P}\in\mathbf{P}}(1)$. Thus by Equation (B.2.7), $\dddot{M}_T(\tilde{\theta}_T) = O_{\mathcal{P}\in\mathbf{P}}(1)$.
- By Equation (B.2.8), the minimum eigenvalue of $-\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$ is bounded above that of positive definite matrix H. Thus $\|\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\| = O_{\mathcal{P}\in\mathbf{P}}(1)$.
- By Condition 4.3.5, $\sup_{\mathcal{P}\in\mathbf{P}} \|\Sigma_T(\mathcal{P})^{1/2}\| = O(1)$.

Thus, by Slutsky's Theorem and Equation (B.2.10), we have that

$$\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \quad \text{(B.2.11)}$$

Lastly, to show our desired result, that $\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d)$ uniformly over $\mathcal{P} \in \mathbf{P}$, by Equation (B.2.11) and Slutsky's Theorem it is sufficient to show that $\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\Sigma_T(\mathcal{P})^{1/2} \xrightarrow{P} I_d$ uniformly over $\mathcal{P} \in \mathbf{P}$. Note if we can show that $\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1} \xrightarrow{P} I_d$ uniformly over $\mathcal{P} \in \mathbf{P}$, then $\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\Sigma_T(\mathcal{P})^{1/2} = \Sigma_T(\mathcal{P})^{-1/2}\left[I_d + o_{\mathcal{P}\in\mathbf{P}}(1)\right]\Sigma_T(\mathcal{P})^{1/2} = I_d + \Sigma_T(\mathcal{P})^{-1/2}o_{\mathcal{P}\in\mathbf{P}}(1)\Sigma_T(\mathcal{P})^{1/2} = I_d + o_{\mathcal{P}\in\mathbf{P}}(1)$. The last limit holds since $\|\Sigma_T(\mathcal{P})^{-1/2}\| =$

$O_{\mathcal{P}\in\mathbf{P}}(1)$ and $\|\Sigma_T(\mathcal{P})^{1/2}\| = O_{\mathcal{P}\in\mathbf{P}}(1)$ by Condition 4.3.5 (use the same argument as that used in the bullet points below Equation (B.2.10)).

Thus it is sufficient to show that $\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1} \xrightarrow{p} I_d$ uniformly over $\mathcal{P} \in \mathbf{P}$. By Taylor's Theorem, for some random $\bar{\theta}_T$ on the line segment between $\hat{\theta}_T$ and $\theta^*(\mathcal{P})$,

$$\ddot{M}_T(\hat{\theta}_T) = \ddot{M}_T(\theta^*(\mathcal{P})) + \dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T - \theta^*(\mathcal{P})).$$

Recall that the probability the inverse of $\ddot{M}_T(\theta^*(\mathcal{P}))$ exists goes to 1 by Equation (B.2.8) (use the same argument as that used in the bullet points below Equation (B.2.10)). Thus we have that $\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$ equals the following:

$$\left[\ddot{M}_T(\theta^*(\mathcal{P})) + \dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T - \theta^*(\mathcal{P}))\right]\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$$

$$= I_d + \dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T - \theta^*(\mathcal{P}))\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$$

Note that $\dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T - \theta^*(\mathcal{P}))\ddot{M}_T(\theta^*(\mathcal{P}))^{-1} = o_{\mathcal{P}\in\mathbf{P}}(1)$ because

- By uniform consistency of $\hat{\theta}_T$, $1_{\|\tilde{\theta}_T - \theta^*(\mathcal{P})\| \leq \varepsilon} \dddot{m} = o_{\mathcal{P}\in\mathbf{P}}(1)$. Thus by Equation (B.2.7), $\dddot{M}_T(\tilde{\theta}_T) = O_{\mathcal{P}\in\mathbf{P}}(1)$.
- By uniform consistency of $\hat{\theta}_T$, $\|\hat{\theta}_T - \theta^*(\mathcal{P})\| = o_{\mathcal{P}\in\mathbf{P}}(1)$.
- By Equation (B.2.8), $\|\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\| = O_{\mathcal{P}\in\mathbf{P}}(1)$.

Asymptotic Normality of $\Sigma_T(\mathcal{P})^{-1/2}\sqrt{T}\dot{M}_T(\theta^*(\mathcal{P}))$

We will show that Equation (B.2.6) holds by applying a martingale central limit theorem. For notational convenience, we let $\dot{m}_{\theta,t} := \dot{m}_\theta(Y_t, X_t, A_t)$. Note that by defini-

tion $\Sigma_T(\mathcal{P})^{-1/2}\sqrt{T}\dot{M}_T(\theta^*(\mathcal{P})) = \Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t\dot{m}_{\theta^*(\mathcal{P}),t}$. We first show that $\left\{\Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}W_t\dot{m}_{\theta^*(\mathcal{P}),t}\right\}_{t=1}^{T}$ is a martingale difference sequence with respect to the histories $\{\mathcal{H}_t\}_{t=0}^{T}$. For any $t \in [1\colon T]$,

$$\mathbb{E}_{\mathcal{P},\pi}\left[\frac{1}{\sqrt{T}}\Sigma_T(\mathcal{P})^{-1/2}W_t\mathbf{c}^\top\dot{m}_{\theta^*(\mathcal{P}),t}\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(a)}{=}\frac{1}{\sqrt{T}}\mathbb{E}_{\mathcal{P},\pi}\left[\mathbb{E}_{\mathcal{P}}\left[\Sigma_T(\mathcal{P})^{-1/2}W_t\mathbf{c}^\top\dot{m}_{\theta^*(\mathcal{P}),t}\middle|\mathcal{H}_{t-1},X_t,A_t\right]\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(b)}{=}\frac{1}{\sqrt{T}}\Sigma_T(\mathcal{P})^{-1/2}\mathbb{E}_{\mathcal{P},\pi}\left[W_t\mathbf{c}^\top\mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P}),t}\middle|\mathcal{H}_{t-1},X_t,A_t\right]\middle|\mathcal{H}_{t-1}\right]\underset{(c)}{=}0$$

- Above, (a) holds by law of iterated expectations.
- (b) holds since $W_t \in \sigma(\mathcal{H}_{t-1},X_t,A_t)$ and since $\Sigma_T(\mathcal{P})$ are a function of stabilizing policies $\{\pi_t^{\text{sta}}\}_{t\geq 1}$, which are pre-specified.
- By Condition 4.3.1, $\mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P}),t}\middle|\mathcal{H}_{t-1},X_t,A_t\right] = \mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P}),t}\middle|X_t,A_t\right]$. Equality (c) holds because $\mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P}),t}\middle|X_t,A_t\right] = 0$ with probability 1 by Condition 4.3.7; note that $\theta^*(\mathcal{P})$ is a critical point of $\mathbb{E}_{\mathcal{P}}[m_{\theta,t}|X_t,A_t]$.

By Cramer-Wold device, to show that Equation (B.2.6) holds, it is sufficient to show that for any fixed $\mathbf{c} \in \mathbb{R}^d$ with $\|\mathbf{c}\|_2 = 1$, that $\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t\dot{m}_{\theta^*(\mathcal{P}),t} \xrightarrow{D} \mathcal{N}\left(0,\mathbf{c}^\top I_d\mathbf{c}\right)$ uniformly over $\mathcal{P} \in \mathbf{P}$. We now apply Theorem B.2.1, a uniform version of the martingale central limit theorem of Dvoretzky [29]; while the original theorem holds for any fixed $\mathcal{P}$, we can show uniform convergence in distribution by ensuring that the conditions of the theorem hold uniformly over $\mathcal{P} \in \mathbf{P}$ (see Definition B.2.3). By Theorem B.2.1, it is sufficient to show that the following two conditions hold:

1. **Conditional Variance:** $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[\left\{\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}W_t\dot{m}_{\vartheta^*(\mathcal{P}),t}\right\}^2\middle|\mathcal{H}_{t-1}\right]\xrightarrow{P}\sigma^2$ uniformly over $\mathcal{P}\in\mathbf{P}$.

2. **Conditional Lindeberg:** For any $\delta>0$,
$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[\left\{\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}W_t\dot{m}_{\vartheta^*(\mathcal{P}),t}\right\}^2 1_{|\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}W_t\dot{m}_{\vartheta^*(\mathcal{P}),t}|>\delta\sqrt{T}}\middle|\mathcal{H}_{t-1}\right]\xrightarrow{P}0$ uniformly over $\mathcal{P}\in\mathbf{P}$.

1. Conditional Variance

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[\left(\mathbf{c}^\top W_t\Sigma_T(\mathcal{P})^{-1/2}\dot{m}_{\vartheta^*(\mathcal{P}),t}\right)^2\middle|\mathcal{H}_{t-1}\right]$$

$$=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(a)}{=}\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\middle|\mathcal{H}_{t-1}\right]\right\}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}$$

$$\underset{(b)}{=}\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}$$
$$\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}}\pi_t(a,X_t,\mathcal{H}_{t-1})\mathbb{E}_{\mathcal{P}}\left[W_t^2\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\middle|\mathcal{H}_{t-1},X_t,A_t=a\right]da\middle|\mathcal{H}_{t-1}\right]\right\}$$
$$\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}$$

$$
\begin{aligned}
&\underset{(c)}{=} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \\
&\qquad \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t^{\mathrm{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}} \left[\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t} \middle| \mathcal{H}_{t-1}, X_t, A_t = a \right] da \middle| \mathcal{H}_{t-1} \right] \right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} \\
&\underset{(d)}{=} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\mathbb{E}_{\mathcal{P}, \pi_t^{\mathrm{sta}}} \left[\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t} \middle| X_t \right] \middle| \mathcal{H}_{t-1} \right] \right\} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} \\
&\underset{(e)}{=} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi_t^{\mathrm{sta}}} \left[\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t} \right] \right\} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} \\
&\underset{(f)}{=} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \Sigma_T(P) \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} = \mathbf{c}^\top I_d \mathbf{c}
\end{aligned}
$$

- Above, (a) holds since $\Sigma_T(\mathcal{P})$ are a function of stabilizing policies $\{\pi_t^{\mathrm{sta}}\}_{t \geq 1}$, which are pre-specified.
- Equality (b) holds by law of iterated expectations.
- Equality (c) holds since $W_t = \sqrt{\frac{\pi_t^{\mathrm{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}} \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$.
- Equality (d) holds because by Condition 4.3.1, $\mathbb{E}_{\mathcal{P}}[\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1}, X_t, A_t = a] = \mathbb{E}_{\mathcal{P}}[\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t} | X_t, A_t = a]$ and by law of iterated expectations.
- Equality (e) holds because by Condition 4.3.1, the distribution of X_t does not depend on $\mathcal{H}_{t-1}$, so $\mathbb{E}_{\mathcal{P}} \left[\mathbb{E}_{\mathcal{P}, \pi_t^{\mathrm{sta}}} \left[\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t} \middle| X_t \right] \middle| \mathcal{H}_{t-1} \right] = \mathbb{E}_{\mathcal{P}} \left[\mathbb{E}_{\mathcal{P}, \pi_t^{\mathrm{sta}}} \left[\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t} \middle| X_t \right] \right] = \mathbb{E}_{\mathcal{P}, \pi_t^{\mathrm{sta}}} \left[\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t} \right]$; the last equality holds by law of iterated expectations.
- Equality (f) holds by definition.

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[\left(\mathbf{c}^\top W_t \Sigma_T(\mathcal{P})^{-1/2}\dot{m}_{\vartheta^*(\mathcal{P}),t}\right)^2 1_{\left|\mathbf{c}^\top W_t \Sigma_T(\mathcal{P})^{-1/2}\dot{m}_{\vartheta^*(\mathcal{P}),t}\right|>\delta\sqrt{T}}\middle|\mathcal{H}_{t-1}\right]$$

$$= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}1_{\left|\mathbf{c}^\top W_t \Sigma_T(\mathcal{P})^{-1/2}\dot{m}_{\vartheta^*(\mathcal{P}),t}\right|>\delta\sqrt{T}}\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(a)}{\leq}\frac{1}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[W_t^4\left(\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}\right)^2\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(b)}{\leq}\frac{\rho_{\max}}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2\left(\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}\right)^2\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(c)}{=}\frac{\rho_{\max}}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}}\pi_t(a,X_t,\mathcal{H}_{t-1})\right.$$
$$\left.\mathbb{E}_{\mathcal{P}}\left[W_t^2\left(\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}\right)^2\middle|\mathcal{H}_{t-1},X_t,A_t=a\right]da\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(d)}{=}\frac{\rho_{\max}}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}}\pi_t^{\text{sta}}(a,X_t)\right.$$
$$\left.\mathbb{E}_{\mathcal{P}}\left[\left(\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}\right)^2\middle|\mathcal{H}_{t-1},X_t,A_t=a\right]da\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(e)}{=}\frac{\rho_{\max}}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P}}\left[\left(\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}\right)^2\middle|X_t\right]\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(f)}{=}\frac{\rho_{\max}}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\left(\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}\right)^2\right]\underset{(g)}{\to}0$$

- Above, inequality (a) holds because $1_{\left|W_t \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}_{\vartheta^*(\mathcal{P}),t}\right| > \sqrt{T}\partial} = 1$ if and only if $W_t^2 \frac{1}{T\partial^2} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} > 1$.

- Inequality (b) holds because by Condition 4.3.9, $W_t^2 \leq \rho_{\max}$ with probability 1.

- Equality (c) holds by the law of iterated expectations.

- Equality (d) holds since $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}} \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$.

- Equality (e) holds because by Condition 4.3.1,
  $\mathbb{E}_{\mathcal{P}} \left[(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c})^2 \middle| \mathcal{H}_{t-1}, X_t, A_t = a \right]$
  $= \mathbb{E}_{\mathcal{P}} \left[(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c})^2 \middle| X_t \right]$ and by law of iterated expectations.

- Equality (f) holds since the distribution of X_t does not depend on $\mathcal{H}_{t-1}$ by Condition 4.3.1 and by law of iterated expectations.

- Regarding limit (g), it is sufficient to show that
  $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\left(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} \right)^2 \right]$ is uniformly bounded over $\mathcal{P} \in \mathbf{P}$ for all sufficiently large T. By Condition 4.3.5, the minimum eigenvalue of $\Sigma_T(P)$ is bounded above zero uniformly over $\mathcal{P} \in \mathbf{P}$ for all sufficiently large T; this bounds the maximum eigenvalue of $\Sigma_T(P)^{-1}$. Also by Condition 4.3.5 the fourth moment of $\dot{m}_{\vartheta^*(\mathcal{P}),t}$ with respect to $\mathcal{P}$ and policy π_t^{sta} is uniformly bounded over $\mathcal{P} \in \mathbf{P}$ and $t \geq 1$. With these two properties we have that
  $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\left(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}^{\otimes 2}_{\vartheta^*(\mathcal{P}),t} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} \right)^2 \right]$ is uniformly bounded over $\mathcal{P} \in \mathbf{P}$ for all sufficiently large T.

Showing that $\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|\leq\varepsilon_{\dddot{m}}} \left\|\dddot{M}_T(\theta)\right\|_1$ is bounded in probability

Recall that for any $B \in \mathbb{R}^{d\times d\times d}$, we denote $\|B\|_1 = \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d |B_{i,j,k}|$. We abbreviate $\dddot{m}_\theta(Y_t, X_t, A_t)$ with $\dddot{m}_{\theta,t}$.

By triangle inequality, $\left\|\dddot{M}_T(\theta)\right\|_1 = \left\|\frac{1}{T}\sum_{t=1}^T W_t \dddot{m}_{\theta,t}\right\|_1 \leq \frac{1}{T}\sum_{t=1}^T W_t \left\|\dddot{m}_{\theta,t}\right\|_1$. Thus we have that

$$\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|\leq\varepsilon_{\dddot{m}}} \left\|\dddot{M}_T(\theta)\right\|_1 \leq \sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|\leq\varepsilon_{\dddot{m}}} \frac{1}{T}\sum_{t=1}^T W_t \left\|\dddot{m}_{\theta,t}\right\|_1 .$$

By Condition 4.3.6 (ii), there exists a function $\dddot{m}$ (note it is not indexed by θ) such that for all $\mathcal{P} \in \mathbf{P}$, we have that $\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|\leq\varepsilon_{\dddot{m}}} \left\|\dddot{m}_{\theta,t}\right\|_1 \leq \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1$.

$$\leq \frac{1}{T}\sum_{t=1}^T W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 .$$

Adding and subtracting $\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi}\left[W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 \middle| \mathcal{H}_{t-1}\right]$,

$$= \frac{1}{T}\sum_{t=1}^T W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 - \mathbb{E}_{\mathcal{P},\pi}\left[W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 \middle| \mathcal{H}_{t-1}\right] + \mathbb{E}_{\mathcal{P},\pi}\left[W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 \middle| \mathcal{H}_{t-1}\right] .$$

By second moment bounds on $\left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1$ from Condition 4.3.6 (i), by Lemma B.2.1, we have that $\frac{1}{T}\sum_{t=1}^T W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 - \mathbb{E}_{\mathcal{P},\pi}\left[W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 \middle| \mathcal{H}_{t-1}\right] = o_{\mathcal{P}\in\mathbf{P}}(1)$.

$$= o_{\mathcal{P}\in\mathbf{P}}(1) + \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi}\left[W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 \middle| \mathcal{H}_{t-1}\right]$$

Since by Condition 4.3.9, $\frac{W_t}{\sqrt{\rho_{\min}}} \geq 1$ with probability 1,

$$\leq o_{\mathcal{P}\in\mathbf{P}}(1) + \frac{1}{T\sqrt{\rho_{\min}}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[W_t^2 \left\| \ddot{m}(Y_t, X_t, A_t) \right\|_1 | \mathcal{H}_{t-1} \right]$$

Since $W_t^2 = \frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$ and by Condition 4.3.1,

$$= o_{\mathcal{P}\in\mathbf{P}}(1) + \frac{1}{T\sqrt{\rho_{\min}}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\left\| \ddot{m}(Y_t, X_t, A_t) \right\|_1 \right] = O_{\mathcal{P}\in\mathbf{P}}(1).$$

Note that by Jensen's inequality, $\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\left\| \ddot{m}(Y_t, X_t, A_t) \right\|_1 \right] \leq \sqrt{\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\left\| \ddot{m}(Y_t, X_t, A_t) \right\|_1^2 \right]}$. By Condition 4.3.6 (i), $\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\left\| \ddot{m}(Y_t, X_t, A_t) \right\|_1^2 \right]$ is bounded, which implies the final limit above.

Lower bounding $-\ddot{M}_T(\theta^*(\mathcal{P}))$

We now show that $-\ddot{M}_T(\theta^*(\mathcal{P})) \succeq H + o_{\mathcal{P}\in\mathbf{P}}(1)$, for positive definite matrix H introduced in Condition 4.3.7 (ii).

By Condition 4.3.5 and Lemma B.2.1, $\frac{1}{T}\sum_{t=1}^{T} W_t \ddot{m}_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi} \left[W_t \ddot{m}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1} \right] = o_{\mathcal{P}\in\mathbf{P}}(1)$, so

$$-\ddot{M}_T(\theta^*(\mathcal{P})) = -\frac{1}{T}\sum_{t=1}^{T} W_t \ddot{m}_{\theta^*(\mathcal{P}),t} = o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[W_t \ddot{m}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1} \right]$$

By law of iterated expectations,

$$= o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[W_t \mathbb{E}_{\mathcal{P}} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1}, X_t, A_t \right] | \mathcal{H}_{t-1} \right]$$

By Condition 4.3.1,

$$= o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}\left[W_t \mathbb{E}_{\mathcal{P}}\left[\ddot{m}_{\theta^*(\mathcal{P}),t}|X_t, A_t\right]|\mathcal{H}_{t-1}\right]$$

By Condition 4.3.7, we have that $\mathbb{E}_{\mathcal{P}}\left[\ddot{m}_{\theta^*(\mathcal{P}),t}|X_t, A_t\right] \preceq 0$; recall that $\theta^*(\mathcal{P})$ is a maximizing value of $\mathbb{E}_{\mathcal{P},\pi}\left[m_{\theta,t}|X_t, A_t\right]$. Also since $\frac{W_t}{\sqrt{\rho_{\max}}} \leq 1$ with probability 1 by Condition 4.3.9,

$$\succeq o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 \mathbb{E}_{\mathcal{P},\pi}\left[\ddot{m}_{\theta^*(\mathcal{P}),t}|X_t, A_t\right]|\mathcal{H}_{t-1}\right]$$

Since $W_t^2 = \frac{\pi_t^{\text{sta}}(A_t,X_t)}{\pi_t(A_t,X_t,\mathcal{H}_{t-1})}$,

$$= o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\ddot{m}_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}\right]$$

Note for any $t \geq 1$, $\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\ddot{m}_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}\right] = \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\ddot{m}_{\theta^*(\mathcal{P}),t}\right]$ because $\{\pi_t^{\text{sta}}\}_{t\geq 1}$ are prespecified. Recall by Condition 4.3.7 $-\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\ddot{m}_{\theta^*(\mathcal{P}),t}\right] \succeq H$ for all sufficiently large T and all $\mathcal{P} \in \mathbf{P}$. Thus our final result is that

$$-\ddot{M}_T(\theta^*(\mathcal{P})) \succeq H + o_{\mathcal{P}\in\mathbf{P}}(1). \tag{B.2.12}$$

B.2.4 Lemmas and Other Helpful Results

Theorem B.2.1 (Uniform Martingale Central Limit Theorem)**.** *Let $\{Z_T(\mathcal{P})\}_{T\geq 1}$ be a sequence of random variables whose distributions are defined by some $\mathcal{P} \in \boldsymbol{P}$ and some nuisance component η. Moreover, let $\{Z_T(\mathcal{P})\}_{T\geq 1}$ be a martingale difference sequence with respect to $\mathcal{F}_t$, meaning $\mathbb{E}_{\mathcal{P},\eta}[Z_t(\mathcal{P})|\mathcal{F}_{t-1}] = 0$ for all $t \geq 1$ and $\mathcal{P} \in \boldsymbol{P}$.*

(a) $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\eta}[Z_t(\mathcal{P})^2|\mathcal{F}_{t-1}] \xrightarrow{P} \sigma^2$ *uniformly over* $\mathcal{P} \in \boldsymbol{P}$*, where* σ^2 *is a constant* $0 < \sigma^2 < \infty$.

(b) For any $\varepsilon > 0$, $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\eta}[Z_t(\mathcal{P})^2 1_{|Z_t(\mathcal{P})|>\varepsilon}|\mathcal{F}_{t-1}] \xrightarrow{P} 0$ *uniformly over* $\mathcal{P} \in \boldsymbol{P}$.

Under the above conditions,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t(\mathcal{P}) \xrightarrow{D} \mathcal{N}(0,\sigma^2) \text{ uniformly over } \mathcal{P} \in \boldsymbol{P}.$$

Proof: By by Kasy [55], Lemma 1, it is sufficient to show that for any sequence $\{\mathcal{P}_T\}_{T=1}^{\infty}$ with $\mathcal{P}_T \in \mathbf{P}$ for all $T \geq 1$, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t(\mathcal{P}_T) \xrightarrow{D} \mathcal{N}(0,\sigma^2)$. In this setting, since $\mathcal{P}_T$ depends on T, we consider triangular array asymptotics and additionally index by T, e.g., $\mathcal{F}_{T,t}$.

Note that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P}_T,\eta}[Z_t(\mathcal{P}_T)^2|\mathcal{F}_{T,t-1}] \xrightarrow{P} \sigma^2$, by Kasy [55], Lemma 1 and condition (a) above.

Also, for any $\varepsilon > 0$, $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P}_T,\eta}\left[Z_t(\mathcal{P}_T)^2 1_{|Z_t(\mathcal{P}_T)|>\varepsilon}\middle|\mathcal{F}_{T,t-1}\right] \xrightarrow{P} 0$, by Kasy [55], Lemma 1 and condition (b) above.

Thus by the martingale central limit theorem of Dvoretzky [29], we have that for the sequence $\{\mathcal{P}_T\}_{T=1}^{\infty}$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t(\mathcal{P}_T) \xrightarrow{D} \mathcal{N}(0,1).$$

Since the sequence $\{\mathcal{P}_T\}_{T=1}^{\infty}$ were chosen arbitrarily from $\mathbf{P}$, the desired result is implied again by Kasy [55], Lemma 1.

Lemma B.2.1. *Let* $f(Y_t, X_t, A_t) \in \mathbb{R}^{d_f}$ *be a function such that*

$\sup_{\mathcal{P}\in\boldsymbol{P},t\geq 1}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\|f(Y_t,X_t,A_t)\|^2\right] < m$ *for some* $m < \infty$*. Under Conditions 4.3.1 and*

4.3.9,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{W_t f(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}[W_t f(Y_t, X_t, A_t)|\mathcal{H}_{t-1}]\right\} = O_{\mathcal{P}\in\boldsymbol{P}}(1). \tag{B.2.13}$$

Note that the above equation implies that

$$\frac{1}{T}\sum_{t=1}^{T}\left\{W_t f(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}[W_t f(Y_t, X_t, A_t)|\mathcal{H}_{t-1}]\right\} = o_{\mathcal{P}\in\boldsymbol{P}}(1).$$

Lemma B.2.1 is a type of martingale weak law of large number result and the proof is similar to the weak law of large numbers proofs for i.i.d. random variables.

Proof: We denote the $k^{\text{th}} \in [1\colon d_f]$ dimension of vector $f(Y_t, X_t, A_t)$ as $f^k(Y_t, X_t, A_t)$. It is sufficient to show the result for any dimension of vector $f(Y_t, X_t, A_t)$. For notational convenience, let $f_t := f^k(Y_t, X_t, A_t)$. Let $\varepsilon > 0$.

$$\begin{aligned}
&\sup_{\mathcal{P}\in\mathbf{P}}\mathbb{P}_{\mathcal{P},\pi}\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t|\mathcal{H}_{t-1}]\right\}\right| > \varepsilon\right)\\
&\underset{(a)}{\leq} \frac{1}{T\varepsilon^2}\sup_{\mathcal{P}\in\mathbf{P}}\mathbb{E}_{\mathcal{P},\pi}\left[\left(\sum_{t=1}^{T}\left\{W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t|\mathcal{H}_{t-1}]\right\}\right)^2\right]\\
&\underset{(b)}{=} \frac{1}{T\varepsilon^2}\sup_{\mathcal{P}\in\mathbf{P}}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[\left\{W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t|\mathcal{H}_{t-1}]\right\}^2\right]\\
&\underset{(c)}{\leq} \frac{1}{T\varepsilon^2}\sup_{\mathcal{P}\in\mathbf{P}}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 f_t^2\right]
\end{aligned}$$

$$\underset{(d)}{=} \frac{1}{T\varepsilon^2} \sup_{\mathcal{P}\in\mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\int_{a\in\mathcal{A}} W_t^2 \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}}[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t = a] da \right]$$

$$\underset{(e)}{=} \frac{1}{T\varepsilon^2} \sup_{\mathcal{P}\in\mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\int_{a\in\mathcal{A}} \pi_t^{\mathrm{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}}[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t = a] da \right]$$

$$\underset{(f)}{=} \frac{1}{T\varepsilon^2} \sup_{\mathcal{P}\in\mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\mathrm{sta}}} \left[f_t^2 \right] \underset{(g)}{\leq} \frac{4m}{\varepsilon^2}$$

- Above (a) holds by Chebyshev's inequality.

- (b) holds because the above terms form a martingale difference sequence with respect to $\mathcal{H}_{t-1}$, i.e., $\mathbb{E}_{\mathcal{P},\pi}\left[W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \middle| \mathcal{H}_{t-1} \right] = 0$; this implies that cross terms disappear, i.e., for $t > s$,

$$\mathbb{E}_{\mathcal{P},\pi}\left[\left(W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \right) \left(W_s f_s - \mathbb{E}_{\mathcal{P},\pi}[W_s f_s | \mathcal{H}_{s-1}] \right) \right]$$

$$= \mathbb{E}_{\mathcal{P},\pi}\left[\mathbb{E}_{\mathcal{P},\pi}\left[\left(W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \right) \left(W_s f_s - \mathbb{E}_{\mathcal{P},\pi}[W_s f_s | \mathcal{H}_{s-1}] \right) \middle| \mathcal{H}_{t-1} \right] \right]$$

Since $s > t$,

$$= \mathbb{E}_{\mathcal{P},\pi}\left[\left(W_s f_s - \mathbb{E}_{\mathcal{P},\pi}[W_s f_s | \mathcal{H}_{s-1}] \right) \mathbb{E}_{\mathcal{P},\pi}\left[W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \middle| \mathcal{H}_{t-1} \right] \right] = 0.$$

- (c) holds because $\mathbb{E}_{\mathcal{P},\pi}\left[\{ W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \}^2 \right]$

$= \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 f_t^2 \right] - \mathbb{E}_{\mathcal{P},\pi}\left[\mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}]^2 \right] \leq \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 f_t^2 \right]$.

- (d) holds by law of iterated expectations.

- (e) holds because $W_t = \sqrt{\frac{\pi_t^{\mathrm{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$.

- (f) holds since by Condition 4.3.1, $\mathbb{E}_{\mathcal{P}}[f_t^2|\mathcal{H}_{t-1}, X_t, A_t] = \mathbb{E}_{\mathcal{P}}[f_t^2|X_t, A_t]$ and by law of iterated expectations $\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[f_t^2] = \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t^{\text{sta}}(a, X_t)\mathbb{E}_{\mathcal{P}}[f_t^2|X_t, A_t = a]da\right]$.

- (g) holds since $\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[f_t^2] < m < \infty$.

Lemma B.2.2. *Let $m_{\theta,t} := m_\theta(Y_t, X_t, A_t)$. Under Conditions 4.3.1, 4.3.3, 4.3.4, 4.3.5, 4.3.7, and 4.3.9,*

$$\sup_{\theta\in\Theta}\left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\} = O_{\mathcal{P}\in\boldsymbol{P}}(1). \tag{B.2.14}$$

Lemma B.2.1 is a type of martingale functionally uniform law of large number result and the proof is similar to the functionally uniform law of large numbers proofs for i.i.d. random variables Van Der Vaart & Wellner [100], Theorem 2.4.1.

Proof:

Finite Bracketing Number: Let $\delta > 0$. We construct a set B_δ which is made up of pairs of functions (l, u). We show that we can find B_δ that satisfies the following:

(a) For any $\theta \in \Theta$, we can find $(l, u) \in B_\delta$ such that

(i) $l(y, x, a) \leq m_\theta(y, x, a) \leq u(y, x, a)$ for all (x, y) in the joint support of $\{\mathcal{P} \in \mathbf{P}\}$ and all $a \in \mathcal{A}$.

(ii) $\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[|u(Y_t, X_t, A_t) - l(Y_t, X_t, A_t)|] \leq \delta$.

(b) The number of pairs in this set is finite, i.e., $|B_\delta| < \infty$.

(c) For any $(l, u) \in B_\delta$, for some $m < \infty$ which does no depend on δ,

$\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[u(Y_t, X_t, A_t)^2] \leq m$ and $\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[l(Y_t, X_t, A_t)^2] \leq m$.

Showing that we can find B_δ that satisfy (a), means that $|B_\delta|$ is an upper bound on the bracketing number of $\{m_\theta : \theta \in \Theta\}$. For more information on bracketing functions, see Van Der Vaart & Wellner [100] and Van der Vaart [99].

To construct B_δ, we follow a similar argument to Example 19.7 of Van der Vaart [99] (page 271). Make a grid over Θ with meshwidth $\lambda/2 > 0$ and let the points in this grid be the set $G_{\lambda/2} \subseteq \Theta$; we will specify λ later. Note that by construction, for any $\theta \in \Theta$ we can find a $\theta \in G_{\lambda/2}$ such that $\|\theta' - \theta\| \leq \lambda$.

By our Lipschitz Condition 4.3.4, we have that for any $\theta, \theta' \in \Theta$, $|m_\theta(Y_t, X_t, A_t) - m_{\theta'}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t)\|\theta - \theta'\|$ for function g such that for some $m_g < \infty$,

$$\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}[g(Y_t, X_t, A_t)^2] \leq m_g. \tag{B.2.15}$$

We now show we can choose $B_\delta = \left\{(m_\theta - g(Y_t, X_t, A_t), m_\theta + g(Y_t, X_t, A_t)) : \theta \in G_{\lambda/2}\right\}$. Note that by compactness of Θ, Condition 4.3.3, the number of points in $G_{\lambda/2}$ is finite, so (b) above holds.

To show that (a) holds for our choice of B_δ, recall that for any $\theta \in \Theta$ we can find a $\theta' \in G_{\lambda/2}$ such that $\|\theta' - \theta\| \leq \lambda$. Also, by the Lipschitz Condition 4.3.4, $|m_\theta(Y_t, X_t, A_t) - m_{\theta'}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t)\|\theta - \theta'\| \leq g(Y_t, X_t, A_t)\lambda$. Thus we have that

$$m_{\theta'}(Y_t, X_t, A_t) - g(Y_t, X_t, A_t)\lambda \leq m_\theta(Y_t, X_t, A_t) \leq m_{\theta'}(Y_t, X_t, A_t) + g(Y_t, X_t, A_t)\lambda.$$

Note that

$$\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[m_{\theta'}(Y_t, X_t, A_t) + g(Y_t, X_t, A_t)\lambda - \{m_{\theta'}(Y_t, X_t, A_t) - g(Y_t, X_t, A_t)\lambda\}\right]$$

$$= 2\lambda \sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[g(Y_t,X_t,A_t)\right] \leq 2\lambda\sqrt{m_g} < \infty.$$

The inequalities above hold by Equation (B.2.15) and since by Jensen's inequality we have that $\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[g(Y_t,X_t,A_t)\right] \leq \sqrt{\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[g(Y_t,X_t,A_t)^2\right]}$. (a) above holds for our choice of B_δ by letting meshwidth $\lambda = \delta/(2\sqrt{m_g})$.

We now show that (c) above holds. Note that

$$\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\left\{m_\theta(Y_t,X_t,A_t) + g(Y_t,X_t,A_t)\right\}^2\right]$$

$$\leq 3 \sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[m_\theta(Y_t,X_t,A_t)^2\right] + 3 \sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[g(Y_t,X_t,A_t)^2\right]. \tag{B.2.16}$$

Note that the above upper bound, Equation (B.2.16), also holds for

$\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\left\{m_\theta(Y_t,X_t,A_t) - g(Y_t,X_t,A_t)\right\}^2\right]$.

Since, $m_\theta(Y_t,X_t,A_t) = m_\theta(Y_t,X_t,A_t) - m_{\theta^*(\mathcal{P})}(Y_t,X_t,A_t) + m_{\theta^*(\mathcal{P})}(Y_t,X_t,A_t)$,

$$\leq 9 \sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\left\{m_\theta(Y_t,X_t,A_t) - m_{\theta^*(\mathcal{P})}(Y_t,X_t,A_t)\right\}^2\right]$$
$$+ 9 \sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[m_{\theta^*(\mathcal{P})}(Y_t,X_t,A_t)^2\right]$$
$$+ 3 \sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[g(Y_t,X_t,A_t)^2\right].$$

Note that $\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[m_{\theta^*(\mathcal{P})}(Y_t,X_t,A_t)^2\right]$ is bounded by our moment Condition 4.3.5 and that $\sup_{\mathcal{P}\in\mathbf{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[g(Y_t,X_t,A_t)^2\right]$ is bounded by Equation (B.2.15).

By our Lipschitz Condition 4.3.4, for any $\theta \in \Theta$, $|m_\theta(Y_t,X_t,A_t) - m_{\theta^*(\mathcal{P})}(Y_t,X_t,A_t)| \leq$

$g(Y_t, X_t, A_t)\|\theta - \theta^*(\mathcal{P})\|$. Thus,

$$\sup_{\mathcal{P}\in\mathbf{P}, t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\left\{m_\theta(Y_t, X_t, A_t) - m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\right\}^2\right]$$
$$\leq \sup_{\mathcal{P}\in\mathbf{P}, t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[g(Y_t, X_t, A_t)^2\right] \|\theta - \theta^*(\mathcal{P})\|^2.$$

The above is bounded by Equation (B.2.15) and by compactness of Θ, Condition 4.3.3. Thus (c) above holds for our choice of B_δ.

Main Argument: We now show that for any $\varepsilon > 0$,

$$\sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\sup_{\theta\in\Theta}\left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\} > \varepsilon\right) \to 0. \tag{B.2.17}$$

An analogous argument can be made to show that

$\sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\sup_{\theta\in\Theta}\left\{-\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\} > \varepsilon\right) \to 0.$

Let $\delta > 0$; we will choose δ later. Let B_δ be the set of pairs of functions as constructed earlier.

$$\sup_{\theta\in\Theta}\left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\}$$

Note that by (a), we get the following upper bound:

$$\leq \max_{(l,u)\in B_\delta}\left\{\frac{1}{T}\sum_{t=1}^{T} W_t u(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}[W_t l(Y_t, X_t, A_t)|\mathcal{H}_{t-1}]\right\}.$$

By adding and subtracting $\mathbb{E}_{\mathcal{P},\pi}\left[W_t u(Y_t, X_t, A_t)\middle|\mathcal{H}_{t-1}\right]$ and triangle inequality,

$$\leq \max_{(l,u)\in B_\delta}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}\left[W_t\left\{u(Y_t,X_t,A_t)-l(Y_t,X_t,A_t)\right\}\middle|\mathcal{H}_{t-1}\right]\right\}$$
$$+\max_{(l,u)\in B_\delta}\left\{\frac{1}{T}\sum_{t=1}^{T}W_t u(Y_t,X_t,A_t)-\mathbb{E}_{\mathcal{P},\pi}\left[W_t u(Y_t,X_t,A_t)\middle|\mathcal{H}_{t-1}\right]\right\}.$$

Note that by Condition 4.3.9, $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t,X_t)}{\pi_t(A_t,X_t,\mathcal{H}_{t-1})}} \leq \sqrt{\rho_{\max}}$ with probability 1, so

$\mathbb{E}_{\mathcal{P},\pi}\left[W_t\left\{u(Y_t,X_t,A_t)-l(Y_t,X_t,A_t)\right\}\middle|\mathcal{H}_{t-1}\right]$

$\leq \frac{1}{\sqrt{\rho_{\max}}}\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2\left\{u(Y_t,X_t,A_t)-l(Y_t,X_t,A_t)\right\}\middle|\mathcal{H}_{t-1}\right]$

$= \frac{1}{\sqrt{\rho_{\max}}}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[u(Y_t,X_t,A_t)-l(Y_t,X_t,A_t)\right] \leq \frac{1}{\sqrt{\rho_{\max}}}\delta$; the last equality holds by Condition 4.3.1 and the last inequality holds by (a). And since $\max_{i\in[1:\,n]}\{a_i\} \leq \sum_{i=1}^{n}|a_i|$,

$$\leq \frac{1}{\sqrt{\rho_{\max}}}\delta + \sum_{(l,u)\in B_\delta}\left|\frac{1}{T}\sum_{t=1}^{T}W_t u(Y_t,X_t,A_t)-\mathbb{E}_{\mathcal{P},\pi}\left[W_t u(Y_t,X_t,A_t)|\mathcal{H}_{t-1}\right]\right|$$

By Lemma B.2.1 and (c), for any $(l,u)\in B_\delta$,

$\frac{1}{T}\sum_{t=1}^{T}W_t u(Y_t,X_t,A_t)-\mathbb{E}_{\mathcal{P},\pi}\left[W_t u(Y_t,X_t,A_t)\middle|\mathcal{H}_{t-1}\right] = o_{\mathcal{P}\in\mathbf{P}}(1)$. Since $|B_\delta|<\infty$ by (b), the convergence holds for all $(l,u)\in B_\delta$ simultaneously, so

$$= \frac{1}{\sqrt{\rho_{\max}}}\delta + o_{\mathcal{P}\in\mathbf{P}}(1).$$

Equation (B.2.17) holds by choosing $\delta = \sqrt{\rho_{\max}}\varepsilon/2$.

B.2.5 Least-Squares Estimator

We use $\varphi(X_t, A_t)$ to denote a feature vector that constructed using context X_t and action A_t.

Condition B.2.1 (Linear Expected Outcome). *For all $\mathcal{P} \in \boldsymbol{P}$, the following holds w.p.* 1*,*

$$\mathbb{E}_{\mathcal{P}}\left[Y_t | X_t, A_t\right] = \varphi(X_t, A_t)^\top \theta^*(\mathcal{P}).$$

Condition B.2.2 (Moment Conditions for Least Squares). *The fourth moments of*

$\varphi(X_t, A_t)\left(Y_t - \varphi(X_t, A_t)^\top \theta^*(\mathcal{P})\right)$ *and* $\varphi(X_t, A_t)$ *with respect to* $\mathcal{P}$ *and policy* π_t^{sta} *are respectively bounded uniformly over* $\mathcal{P} \in \boldsymbol{P}$ *and* $t \geq 1$*.*

Also the minimum eigenvalue of

$$\Sigma_T(\mathcal{P}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[\varphi(Y_t, X_t, A_t)^{\otimes 2}\left(Y_t - \varphi(Y_t, X_t, A_t)^\top \theta^*(\mathcal{P})\right)^2\right]$$

and $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[\varphi(X_t, A_t)^{\otimes 2}\right]$ *respectively are both bounded above constant some constant greater than zero for all* $\mathcal{P} \in \boldsymbol{P}$*.*

Condition B.2.3 (Importance Ratios for Least Squares). *Let* $\rho_{\min} > 0$ *and* $\rho_{\max,T} > 0$ *be a non-random sequence such that* $\frac{\rho_{\max,T}}{T} \to 0$*.* $\{\pi_t^{\text{sta}}\}_{t=1}^T$ *are pre-specified and do not depend on data* $\{Y_t, X_t, A_t\}_{t=1}^T$*. For all* $\mathcal{P} \in \boldsymbol{P}$*, the following holds w.p.* 1*,*

$$\rho_{\min} \leq \frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})} \leq \rho_{\max,T}.$$

Note that Condition B.2.3 allows $\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ to go to zero at some rate for stabilizing policies $\{\pi_t^{\text{sta}}\}_{t \geq 1}$ that are strictly bounded away from 0 and 1.

We now define the AW-LS estimator for $\theta^*(\mathcal{P}) \in \mathbb{R}^d$:

$$\hat{\theta}_T^{\text{AW-LS}} := \underset{\theta \in \mathbb{R}^d}{\text{argmax}} \left\{ -\sum_{t=1}^{T} W_t \left(Y_t - \varphi(X_t, A_t)^\top \theta\right)^2 \right\}. \tag{B.2.18}$$

Theorem B.2.2 (Consistency and Asymptotic Normality of Adaptively-Weighted Least Squares Estimator)**.** *Under Conditions 4.3.1, B.2.1, B.2.2, and B.2.3,*

$$\Sigma_T(\mathcal{P})^{-1/2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t\varphi(X_t,A_t)^{\otimes 2}\right)\left(\hat{\theta}_T^{\text{AW-LS}}-\theta^*(\mathcal{P})\right)\xrightarrow{D}\mathcal{N}(0,I_d)$$

uniformly over $\mathcal{P}\in\boldsymbol{P}$,

where $\Sigma_T(\mathcal{P}) := \frac{1}{T}\sum_{t=1}^{T}\varphi(X_t,A_t)^{\otimes 2}\left(Y_t-\varphi(X_t,A_t)^\top\theta^*(\mathcal{P})\right)^2$.

Proof: By taking the derivative of Equation (B.2.18) with respect to the parameters, we have that

$$0=\sum_{t=1}^{T} W_t\varphi(X_t,A_t)\left(Y_t-\varphi(X_t,A_t)^\top\hat{\theta}_T^{\text{AW-LS}}\right).$$

By rearranging terms, we have that

$$-\frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t\varphi(X_t,A_t)\left(Y_t-\varphi(X_t,A_t)^\top\theta^*(\mathcal{P})\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t\varphi(X_t,A_t)^{\otimes 2}\left(\hat{\theta}_T^{\text{AW-LS}}-\theta^*(\mathcal{P})\right). \quad \text{(B.2.19)}$$

We first show that the following holds:

$$\Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t\varphi(X_t,A_t)\left(Y_t-\varphi(X_t,A_t)^\top\theta^*(\mathcal{P})\right)\xrightarrow{D}\mathcal{N}(0,I_d)$$

uniformly over $\mathcal{P}\in\mathbf{P}$. (B.2.20)

Equation (B.2.20) holds by a similar argument as that used in Section B.2.3 for

$\dot{m}_\theta(Y_t, X_t, A_t) = \varphi(X_t, A_t)\left(Y_t - \varphi(X_t, A_t)^\top \theta^*(\mathcal{P})\right)$ by showing that the conditions of Theorem B.2.1 hold. It can be checked that all the arguments hold even when we allow $\rho_{\max,T}$ to grow at a rate such that $\frac{\rho_{\max,T}}{T} \to 0$.

By Equations (B.2.19) and (B.2.20),

$$\Sigma_T(\mathcal{P})^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \varphi(X_t, A_t)^{\otimes 2} \left(\hat{\theta}_T^{\text{AW-LS}} - \theta^*(\mathcal{P})\right) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{B.2.21}$$

By Equation (B.2.21), to ensure that $\hat{\theta}_T^{\text{AW-LS}} \xrightarrow{P} \theta^*(\mathcal{P})$ uniformly over $\mathcal{P} \in \mathbf{P}$, it is sufficient to show that the minimum eigenvalue of $\Sigma_T(\mathcal{P})^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \varphi(X_t, A_t)^{\otimes 2}$ goes to infinity uniformly over $\mathcal{P} \in \mathbf{P}$ as $T \to \infty$.

By Condition B.2.2, the maximum eigenvalue of $\Sigma_T(\mathcal{P})$ is bounded uniformly over $\mathcal{P} \in \mathbf{P}$, so the minimum eigenvalue of $\Sigma_T(\mathcal{P})^{-1/2}$ is bounded uniformly above 0. Thus it is sufficient to show that the minimum eigenvalue of $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \varphi(X_t, A_t)^{\otimes 2}$ goes to infinity uniformly over $\mathcal{P} \in \mathbf{P}$ as $T \to \infty$.

Note that by Lemma B.2.1 and Condition B.2.2,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \varphi(X_t, A_t)^{\otimes 2} - \mathbb{E}_{\mathcal{P},\pi}\left[W_t \varphi(X_t, A_t)^{\otimes 2} \middle| \mathcal{H}_{t-1}\right] = O_{\mathcal{P}\in\mathbf{P}}(1). \tag{B.2.22}$$

Note that by law of iterated expectations,

$$\mathbb{E}_{\mathcal{P},\pi}\left[W_t \varphi(X_t, A_t)^{\otimes 2} \middle| \mathcal{H}_{t-1}\right]$$

$$= \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}}\left[W_t \varphi(X_t, A_t)^{\otimes 2} | \mathcal{H}_{t,1}, X_t, a\right] da \middle| \mathcal{H}_{t-1}\right].$$

By Condition 4.3.1 and since $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$,

$$= \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \sqrt{\frac{\pi_t(a, X_t, \mathcal{H}_{t-1})}{\pi_t^{\text{sta}}(a, X_t)}}\pi_t^{\text{sta}}(a, X_t)\mathbb{E}_{\mathcal{P}}\left[\varphi(X_t, A_t)^{\otimes 2}|X_t, a\right] da \middle| \mathcal{H}_{t-1}\right]$$

Since by Condition B.2.3, $\frac{\pi_t(a, X_t, \mathcal{H}_{t-1})}{\pi_t^{\text{sta}}(a, X_t)} \geq \frac{1}{\sqrt{\rho_{\max,T}}}$ and $\varphi(X_t, A_t)^{\otimes 2} \succeq 0$,

$$\succeq \frac{1}{\sqrt{\rho_{\max,T}}}\mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t^{\text{sta}}(a, X_t)\mathbb{E}_{\mathcal{P}}\left[\varphi(X_t, A_t)^{\otimes 2}|X_t, a\right] da \middle| \mathcal{H}_{t-1}\right].$$

Since π_t^{sta} are pre-specified and since by our i.i.d. potential outcomes assumption (Condition 4.3.1) X_t do not depend on $\mathcal{H}_{t-1}$,

$$= \frac{1}{\sqrt{\rho_{\max,T}}}\mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t^{\text{sta}}(a, X_t)\mathbb{E}_{\mathcal{P}}\left[\varphi(X_t, A_t)^{\otimes 2}|X_t, a\right] da\right].$$

By law of iterated expectations,

$$= \frac{1}{\sqrt{\rho_{\max,T}}}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\varphi(X_t, A_t)^{\otimes 2}\right].$$

The above result and Equation (B.2.22) implies that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t\varphi(X_t, A_t)^{\otimes 2} \succeq O_{\mathcal{P}\in\mathbf{P}}(1) + \sqrt{\frac{T}{\rho_{\max,T}}}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\varphi(X_t, A_t)^{\otimes 2}\right]. \quad \text{(B.2.23)}$$

By Condition B.2.2, the minimum eigenvalue of $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[\varphi(X_t, A_t)^{\otimes 2}]$ is bounded above some constant greater than zero for all $\mathcal{P} \in \mathbf{P}$. By Condition B.2.3, $\sqrt{\frac{T}{\rho_{\max,T}}} \to \infty$. Thus by Equation (B.2.21) and Equation (B.2.23), we have that $\hat{\theta}_T^{\text{AW-LS}} \xrightarrow{P} \theta^*(\mathcal{P})$ uniformly

over $\mathcal{P} \in \mathbf{P}$.

B.3 Choice of Stabilizing Policy

B.3.1 Optimal Stabilizing Policy in Multi-Arm Bandit Setting

Here we consider the multi-armed bandit setting where $\mathbb{E}_{\mathcal{P}}[Y_t(a)] = \theta_a^*(\mathcal{P})$ and $\mathrm{Var}_{\mathcal{P}}(Y_t(a)) = \sigma^2$. We consider the adaptively-weighted least-squares estimator where $m_\theta(Y_t, A_t) = -1_{A_t=a}(Y_t - \theta_a^*(\mathcal{P}))^2$. By Theorem 4.3.1, we have that

$$\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[1_{A_t=a}(Y_t-\theta_a^*(\mathcal{P}))^2\right]\right)^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}W_t 1_{A_t=a}\right)\sqrt{T}(\hat{\theta}_{T,a}^{\text{AW-LS}}-\theta_a^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0,1).$$

While the asymptotic variance of $\sqrt{T}(\hat{\theta}_{T,a}^{\text{AW-LS}} - \theta_a^*(\mathcal{P}))$ does not necessarily concentrate we can examine the following:

$$\left(\frac{1}{T}\sum_{t=1}^{T}W_t 1_{A_t=a}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[1_{A_t=a}(Y_t-\theta_a^*(\mathcal{P}))^2\right]\right)\left(\frac{1}{T}\sum_{t=1}^{T}W_t 1_{A_t=a}\right)^{-1}$$

By Lemma B.2.1, we have that $\frac{1}{T}\sum_{t=1}^{T} W_t 1_{A_t=a} - \sqrt{\pi_t^{\text{sta}}(a)\pi_t(A_t,\mathcal{H}_{t-1})} \xrightarrow{P} 0$. Thus we have

$$= \left(\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\text{sta}}(a)\sigma^2\right)\left(o_p(1)+\frac{1}{T}\sum_{t=1}^{T}\sqrt{\pi_t^{\text{sta}}(a)\pi_t(A_t,\mathcal{H}_{t-1})}\right)^{-2}.$$

As long as $\pi_t^{\text{sta}}(a), \pi_t(A_t,\mathcal{H}_{t-1})$ are bounded away from zero w.p. 1, the $o_p(1)$ term is asymptotically negligible and we can just consider $\left(\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\text{sta}}(a)\sigma^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{\pi_t^{\text{sta}}(a)\pi_t(A_t,\mathcal{H}_{t-1})}\right)^{-2}$.

By Cauchy-Schwartz inequality,

$$\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{\pi_t^{\text{sta}}(a)\pi_t(a,\mathcal{H}_{t-1})}\right)^2 \leq \left(\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\text{sta}}(a)\right)\left(\frac{1}{T}\sum_{t=1}^{T}\pi_t(a,\mathcal{H}_{t-1})\right).$$

Thus, $\frac{1}{\frac{1}{T}\sum_{t=1}^{T}\pi_t(a,\mathcal{H}_{t-1})} \leq \frac{\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\text{sta}}(a)}{\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{\pi_t^{\text{sta}}(a)\pi_t(a,\mathcal{H}_{t-1})}\right)^2}$, so

$$\frac{\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\text{sta}}(a)}{\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{\pi_t(a,\mathcal{H}_{t-1})\pi_t^{\text{sta}}(a)}\right)^2} \geq \frac{1}{\frac{1}{T}\sum_{t=1}^{T}\pi_t(a,\mathcal{H}_{t-1})}.$$

Note that this lower bound is achieved when $\pi_t^{\text{sta}}(a) = \pi_t(a)$. However, since π_t is a function of $\mathcal{H}_{t-1}$ and stabilizing policies $\{\pi_t^{\text{sta}}\}_{t=1}^{T}$ are pre-specified, setting $\pi_t^{\text{sta}}(A_t) = \pi_{t,a}$ is generally an unfeasible choice. Thus we want to choose π_t^{sta} to be as close to π_t as possible, subject to the constraint that the stabilizing policies are pre-specified, i.e., not a function of the data $\{Y_t, X_t, A_t\}_{t\geq 1}$.

B.3.2 Approximating the Optimal Stabilizing Policy

One way to approximately choose the optimal evaluation policy is to select $\pi_t^{\text{sta}}(a,x) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$. Note that $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$ depends on the $\mathcal{P}$, which is unknown. Thus it is natural to choose $\pi_t^{\text{sta}}(a,x)$ to be $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$ weighted by a prior on $\mathcal{P}$. Note that as long as the evaluation policy ensures that weights W_t are bounded, the choice of evaluation policy does not affect the asymptotic validity of the estimator.

In Figure B.4, we display the difference in mean squared error for the AW-LS estimator in a two-armed bandit setting for two different choices of evaluation policy: (1) the uniform evaluation policy which selects actions uniformly from $\mathcal{A}$ and (2) the expected $\pi_t(a,\mathcal{H}_{t-1})$ evaluation policy for which $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,\mathcal{H}_{t-1})]$. We can see in this setting that by setting $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,\mathcal{H}_{t-1})]$ we are able to decrease the mean squared error of the AW-LS estimator compared AW-LS with the uniform evaluation policy. Note

though that in some cases setting $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a, \mathcal{H}_{t-1})]$ is equivalent to choosing the uniform evaluation policy. For example, a two-armed bandit with identical arms so under common bandit algorithms $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a, \mathcal{H}_{t-1})] = 0.5$ for all $t \in [1 \colon T]$, which will make the evaluation policy $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a, \mathcal{H}_{t-1})]$ equivalent to the uniform policy.

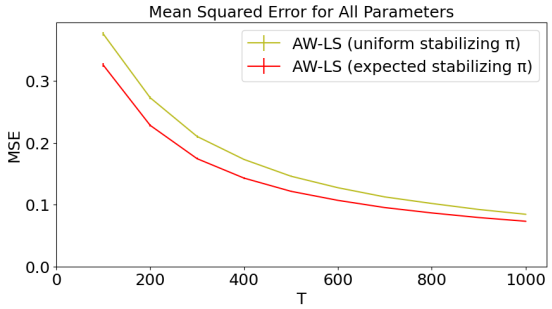

Figure B.4: Above we plot the mean squared errors for the adaptively-weighted least squares estimator with evaluation policies: (1) uniform evaluation policy which selects actions uniformly from $\mathcal{A}$ and (2) expected $\pi_t(a, \mathcal{H}_{t-1})$ evaluation policy for which $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a)]$ (oracle quantity). In a two arm bandit setting we perform Thompson Sampling with standard normal priors, 0.01 clipping, $\theta^*(\mathcal{P}) = [\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})] = [0, 1]$, standard normal errors, and $T = 1000$. Error bars denote standard errors computed over 5,000 Monte Carlo simulations.

B.4 Need for Uniformly Valid Inference on Data Collected with Bandit Algorithms

Here we consider the two-armed bandit setting where $\mathbb{E}_{\mathcal{P}}[R_t(a)] = \theta_{0,a}(\mathcal{P})$, $\text{Var}_{\mathcal{P}}(R_t(a)) = \sigma^2$, and $\mathbb{E}_{\mathcal{P}}[R_t(a)^4] < c < \infty$ for $a \in \{0,1\}$. The unweighted least squares estimator is asymptotically normal on adaptively collected data under the following condition of Lai & Wei[62], there exists a non-random sequence $\{b_t\}_{t\geq 1}$ such that

$$b_T \cdot \sum_{t=1}^{T} A_t \xrightarrow{P} 1. \tag{B.4.1}$$

Specifically, by Theorem 3 of Lai & Wei[62], under (B.4.1),

$$\sqrt{\sum_{t=1}^{T} A_t}(\hat{\theta}^{\text{OLS}}_{T,1} - \theta^*_1(\mathcal{P})) = \frac{\sum_{t=1}^{T} A_t(R_t - \theta^*_1(\mathcal{P}))}{\sqrt{\sum_{t=1}^{T} A_t}} \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

However, as discussed in Deshpande et al.[26] and Zhang et al.[109], (B.4.1) can fail to to hold for common bandit algorithms when there is no unique optimal policy, i.e., when $\theta^*_0(\mathcal{P}) - \theta^*_1(\mathcal{P}) = 0$. For example, in Figure B.5 we plot $\frac{1}{T}\sum_{t=1}^{T} A_t$ for Thompson Sampling and ε-greedy for a bandit with two identical arms.

In order to construct reliable confidence intervals using asymptotic approximations, it is crucial that that estimators converge uniformly in distribution. To illustrate the importance of uniformity, consider the following example. We can modify Thompson Sampling to ensure that $\frac{1}{T}\sum_{t=1}^{T} A_t \xrightarrow{P} 0.5$ when $\theta^*_1(\mathcal{P}) - \theta^*_0(\mathcal{P}) = 0$. For example, we could do this by using an algorithm we call Thompson Sampling Hodges (inspired by the Hodges

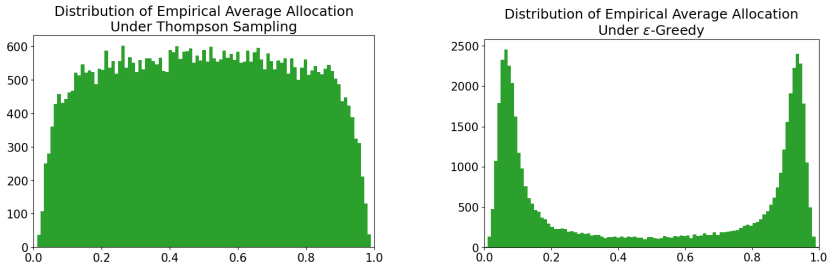

Figure B.5: Above we plot empirical allocations, $\frac{1}{T}\sum_{t=1}^{T} A_t$, under both Thompson Sampling (standard normal priors, 0.01 clipping) and ε-greedy ($\varepsilon = 0.1$) under zero margin $\theta_0^*(\mathcal{P}) = \theta_1^*(\mathcal{P}) = 0$. For our simulations $T = 100$, errors are standard normal, and we use $50k$ Monte Carlo repetitions.

estimator; see Van der Vaart[99], Page 109), defined below:

$$\pi_t(1, \mathcal{H}_{t-1}) = \mathbb{P}(\tilde{\theta}_1 > \tilde{\theta}_0 | \mathcal{H}_{t-1}) 1_{|\mu_{1,t} - \mu_{0,t}| > t^{-4}} + 0.5 1_{|\mu_{1,t} - \mu_{0,t}| \leq t^{-4}}$$

Under standard Thompson Sampling arm one is chosen according to the posterior probability that is optimal, so $\pi_t(1, \mathcal{H}_{t-1}) = \mathbb{P}(\tilde{\theta}_1 > \tilde{\theta}_0 | \mathcal{H}_{t-1})$. Above, $\mu_{a,t}$ denotes the posterior mean for the mean reward for arm a at time t. Under TS-Hodges, if difference between the posterior means, $|\mu_{1,t} - \mu_{0,t}|$, is less than t^{-4}, π_t is set to 0.5. Additionally, we clip the action selection probabilities to bound them strictly away from 0 and 1 for some constant $\pi_{\min}$ in the following sense $\text{clip}(\pi_t) = (1 - \pi_{\min}) \wedge (\pi_t \vee \pi_{\min})$. Under TS-Hodges with clipping, we can show that

$$\frac{1}{T}\sum_{t=1}^{T} A_t \xrightarrow{P} \begin{cases} 1 - \pi_{\min} & \text{if } \theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P}) > 0 \\ \pi_{\min} & \text{if } \theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P}) < 0 \\ 0.5 & \text{if } \theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P}) = 0 \end{cases} \tag{B.4.2}$$

By equation (B.4.2), we satisfy (B.4.1) pointwise for every fixed $\mathcal{P}$ and we have that the

OLS estimator is asymptotically normal *pointwise*[62]. However, equation (B.4.2) fails to hold uniformly over $\mathcal{P} \in \mathbf{P}$. Specifically, it fails to hold for any sequence of $\{\mathcal{P}_t\}_{t=1}^{\infty}$ such that $\theta_1^*(\mathcal{P}_t) - \theta_0^*(\mathcal{P}_t) = t^{-4}$. In Figure B.6, we show that confidence intervals constructed using normal approximations fail to provide reliable confidence intervals, even for very large sample sizes for the worst case values of $\theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P})$.

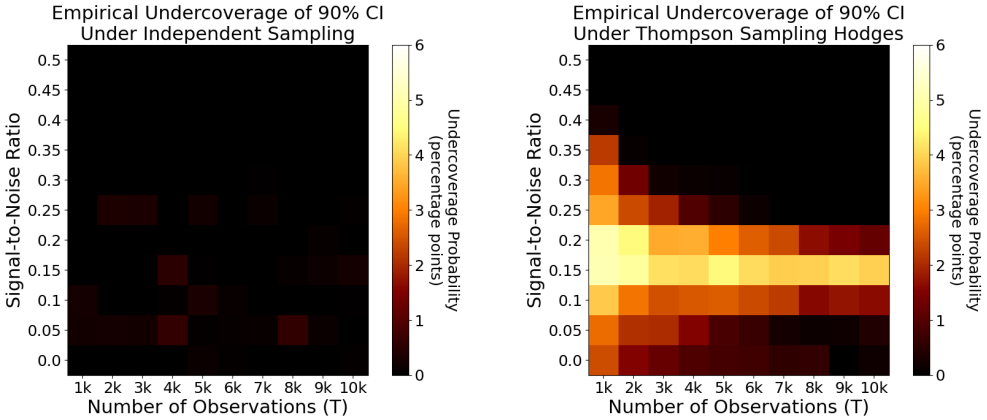

Figure B.6: Above we construct confidence intervals for $\theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P})$ using a normal approximation for the OLS estimator. We compare independent sampling ($\pi_t = 0.5$) and TS Hodges, both with standard normal priors, 0.01 clipping, standard normal errors, and $T = 10,000$. We vary the value of $\theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P})$ in the simulations to demonstrate the non-uniformity of the confidence intervals.

B.5 Discussion of Chen et al., 2020

Here we show formally that Theorem 3.1 in Chen et al. [23], which proves that the OLS estimator is asymptotically normal on data collected with an ε-greedy algorithm, does not cover the case in which there is no unique optimal policy.

They assume that for rewards R_t, context vectors X_t, and binary actions $A_t \in \{0, 1\}$,

$$\mathbb{E}[R_t | X_t, A_t] = A_t \mathbf{X}_t^\top \beta_1 + (1 - A_t)\mathbf{X}_t^\top \beta_0.$$

They define $\beta := \beta_1 - \beta_0$.

Specifically at part 1(b) of their proof on page 4 of the supplementary material, they claim that $g(\hat{\beta}_t, \varepsilon) \xrightarrow{P} g(\beta, \varepsilon)$, where $\hat{\beta}_t$ is the OLS estimator for $\beta := \beta_1 - \beta_0$ and g is defined as follows:

$$g(\beta_0, \beta_1, \varepsilon) = \frac{\varepsilon}{2} \int \mathbf{v}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v} d\mathcal{P}_x + (1 - \varepsilon) \int 1_{\beta^\top \mathbf{x} \geq 0} \mathbf{v}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v} d\mathcal{P}_x$$

Above $v \in \mathbb{R}^d$ is arbitrary fixed vector and $x \in \mathbb{R}^d$ are the context vectors. $\mathcal{P}_x$ is the distribution of the context vectors X_t.

Specifically, they claim that $g(\hat{\beta}_t, \varepsilon) \xrightarrow{P} g(\beta, \varepsilon)$ because $\hat{\beta}_t \xrightarrow{P} \beta$ (Corollary 3.1) and by continuous mapping theorem.

Recall the continuous mapping theorem for convergence in probability[99] Theorem 2.3:

Theorem B.5.1 (Continuous Mapping Theorem). *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$. If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.*

Note that g is not continuous in β at the value $\beta = 0 \in \mathbb{R}^d$; this is due to the indicator

term $1_{\beta^\top \mathbf{x} \geq 0}$. Thus, the standard continuous mapping theorem can not be applied in this setting. Note that the case that $0 = \beta = \beta_1 - \beta_0$, is exactly when there is no unique optimal policy. This means that Theorem 3.1 in Chen et al. [23] does not cover the setting in which there is no unique optimal policy.

C

Inference after Adaptive Sampling for Longitudinal Data

C.1 Examples and Simulation Details

Overview of Appendix C.1.

- **Section C.1.1:** Simulation Details
  - **Section C.1.1:** Additional Information on Reward Generation
  - **Section C.1.1:** Sandwich Variance Estimator
  - **Section C.1.1:** Adaptive Sandwich Variance Estimator
- **Section C.1.2:** Example Algorithms that Satisfy Conditions 5.3.2 and 5.3.3
  - **Section C.1.2:** Boltzmann Exploration Algorithm (Lemma C.1.1)
  - **Section C.1.2:** Stochastic Mirror Descent Algorithm (Lemma C.1.2)
- **Section C.1.3:** Radon-Nikodym Derivatives (Lemma C.1.3)
- **Section C.1.4:** Bracketing Numbers
  - **Section C.1.4:** Definition of Bracketing Numbers
  - **Section C.1.4:** Product of Lipschitz Policy Functions are Lipschitz (Lemma C.1.4)
  - **Section C.1.4:** Bracketing Number for Product of Function Classes (Lemma C.1.5)

C.1.1 Simulation Details

Additional Information on Reward Generation

In the reward generation formula from display (5.6.2), for each user $i \in [1\colon n]$, the errors $\varepsilon_t^{(i)} \sim \mathcal{N}(0,1)$ marginally for each $t \in [1\colon T]$; however, $\mathrm{Corr}(\varepsilon_t^{(i)}, \varepsilon_s^{(i)}) = 0.5^{|t-s|/2}$, which means the reward errors within a user are correlated over time. Additionally, we set the parameters from display (5.6.2) to the following values: $\kappa_0 = 0$, $\kappa_2 = 0$, and we consider simulations with both $\kappa_1 = 1$ and $\kappa_1 = 5$.

Sandwich Variance Estimator

Recall from display (5.5.1) that the sandwich variance is $[\dot{\Psi}^*]^{-1}\Sigma[\dot{\Psi}^*]^{-1,\top}$, where

$$\dot{\Psi}^* \triangleq \frac{\partial}{\partial\theta}\mathbb{E}_{\pi^*_{2:T}}\left[\psi\big(\mathcal{H}_T^{(i)};\theta\big)\right]\Big|_{\theta=\theta^*} \quad \text{and} \quad \Sigma \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\psi\big(\mathcal{H}_T^{(i)};\theta^*\big)^{\otimes 2}\right]. \tag{C.1.1}$$

Above, we use the notation $x^{\otimes 2} \triangleq xx^\top$.

The sandwich variance *estimator* we use is $\underline{\dot{\Psi}}^{-1,\top}\hat{\Sigma}\underline{\dot{\Psi}}^{-1,\top}$, where

$$\underline{\dot{\Psi}} \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\psi\big(\mathcal{H}_T^{(i)};\theta\big)\Big|_{\theta=\hat{\theta}^{(n)}} \quad \text{and} \quad \hat{\Sigma} \triangleq \frac{1}{n}\sum_{i=1}^{n}\psi\big(\mathcal{H}_T^{(i)};\hat{\theta}^{(n)}\big)^{\otimes 2}. \tag{C.1.2}$$

Adaptive Sandwich Variance Estimator

Recall from display (5.5.2) that the sandwich variance is $[\dot{\Psi}^*]^{-1}\Sigma^{\text{adapt}}[\dot{\Psi}^*]^{-1,\top}$, where $\dot{\Psi}^*$ is defined as in display (C.1.1) and

$$\Sigma^{\text{adapt}} \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\left\{\psi(\mathcal{H}_T^{(i)};\theta^*) + \dot{\Psi}^* \sum_{t=1}^{T-1} M_t\, \varphi_t(\mathcal{H}_t^{(i)};\beta_t^*)\right\}^{\otimes 2}\right].$$

By Lemma C.3.1, the adaptive sandwich variance $[\dot{\Psi}^*]^{-1}\Sigma^{\text{adapt}}[\dot{\Psi}^*]^{-1,\top}$ equals the lower-right $\dim(\theta) \times \dim(\theta)$ block of the following matrix:

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} \Sigma_{1:T} \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1,\top} \tag{C.1.3}$$

where

$$\Sigma_{1:T} \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\begin{pmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1^*) \\ \varphi_2(\mathcal{H}_2^{(i)};\beta_2^*) \\ \vdots \\ \varphi_{T-1}(\mathcal{H}_{T-1}^{(i)};\beta_{T-1}^*) \\ \psi(\mathcal{H}_T^{(i)};\theta^*) \end{pmatrix}^{\otimes 2}\right].$$

and

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix} \triangleq \begin{bmatrix} \frac{\partial}{\partial \beta_{1:T-1}} \Phi_{1:T-1}(\beta_{1:T-1}) & \frac{\partial}{\partial \theta} \Phi_{1:T-1}(\beta_{1:T-1}) \\ \frac{\partial}{\partial \beta_{1:T-1}} \Psi(\beta_{1:T-1},\theta) & \frac{\partial}{\partial \theta} \Psi(\beta_{1:T-1},\theta) \end{bmatrix} \Bigg|_{(\beta_{1:T-1},\theta)=(\beta^*_{1:T-1},\theta^*)}.$$

Note above that $\frac{\partial}{\partial\theta}\Phi_{1:T-1}(\beta_{1:T-1}) = 0$ since $\Phi_{1:T-1}(\beta_{1:T-1})$ is not a function of θ. We estimate the entire matrix from display (C.1.3) as follows:

$$\begin{bmatrix} \underline{\dot{\Phi}}_{1:T-1} & 0 \\ \hat{V}_{T,1:T-1} & \underline{\dot{\Psi}} \end{bmatrix}^{-1} \hat{\Sigma}_{1:T} \begin{bmatrix} \underline{\dot{\Phi}}_{1:T-1} & 0 \\ \hat{V}_{T,1:T-1} & \underline{\dot{\Psi}} \end{bmatrix}^{-1,\top}$$

where $\underline{\dot{\Psi}}$ is defined in display (C.1.2),

$$\underline{\dot{\Phi}}_{1:T-1} \triangleq \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta_{1:T-1}} \begin{bmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1) \\ W_2^{(i)}(\beta_1,\hat{\beta}_1^{(n)})\varphi_2(\mathcal{H}_2^{(i)};\beta_2) \\ W_{2:3}^{(i)}(\beta_{1:2},\hat{\beta}_{1:2}^{(n)})\varphi_3(\mathcal{H}_3^{(i)};\beta_3) \\ \vdots \\ W_{2:T-1}^{(i)}(\beta_{1:T-2},\hat{\beta}_{1:T-2}^{(n)})\varphi_{T-1}(\mathcal{H}_{T-1}^{(i)};\beta_{T-1}) \end{bmatrix} \Bigg|_{\beta_{1:T-1}=\hat{\beta}_{1:T-1}^{(n)}},$$

$$\hat{V}_{T,1:T-1} \triangleq \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \beta_{1:T-1}} W_{2:T}^{(i)}(\beta_{1:T-1},\hat{\beta}_{1:T-1}^{(n)}) \right\} \Bigg|_{\beta_{1:T-1}=\hat{\beta}_{1:T-1}^{(n)}} \psi(\mathcal{H}_T^{(i)};\hat{\theta}^{(n)}),$$

and

$$\hat{\Sigma}_{1:T} \triangleq \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \varphi_1(\mathcal{H}_1^{(i)};\hat{\beta}1) \\ \varphi_2(\mathcal{H}_2^{(i)};\hat{\beta}2) \\ \vdots \\ \varphi_{T-1}(\mathcal{H}_{T-1}^{(i)};\hat{\beta}_{T-1}^{(n)}) \\ \psi(\mathcal{H}_T^{(i)};\hat{\theta}^{(n)}) \end{pmatrix}^{\otimes 2}.$$

C.1.2 Example Algorithms that Satisfy Conditions 5.3.2 and 5.3.3

Boltzmann Exploration Algorithm (Lemma C.1.1)

We consider a Boltzmann (or Softmax) exploration type adaptive sampling algorithm[9,20,92] as described in Section 5.6 (Simulation Results). Specifically, we consider a binary action setting ($\mathcal{A} = \{0, 1\}$) and a Boltzmann sampling algorithm that forms action selection probabilities as follows:

$$\pi_t\big(1, S_t^{(i)}; \beta_{t-1}\big) = \text{Clip}_{\pi_{\min}}\Big[\text{expit}\big(\rho \cdot \beta_{t-1,1}^{\top} S_t^{(i)}\big)\Big], \tag{C.1.4}$$

where $\text{Clip}_{\pi_{\min}}[x] \triangleq \min(\max(x, \pi_{\min}), 1 - \pi_{\min})$. Above the parameter ρ is a positive constant that controls the steepness of the Softmax function; larger values of ρ make the Softmax function steeper. The policy parameters $\hat{\beta}_t^{(n)} = \big[\hat{\beta}_{t,0}^{(n)}, \hat{\beta}_{t,1}^{(n)}\big]$ are those from the least squares example defined earlier in display (5.3.8).

Note that exploration Condition 5.3.2 is satisfied because the action selection probabilities are constrained between $[\pi_{\min}, 1 - \pi_{\min}]$. In Lemma C.1.1 below, we show that Condition 5.3.3 holds under the assumption that $\mathbb{E}_{\pi^*_{2:t}}\big[\big\|S_t^{(i)}\big\|_2^{2+\alpha}\big] < \infty$ for all $t \in [2\colon T]$.

Lemma C.1.1 (Boltzmann Exploration Algorithm)**.** *We consider the Boltzmann algorithm example that selects actions as described in display* (C.1.4)*. We show that Condition 5.3.3 holds under the condition that* $\mathbb{E}_{\pi^*_{2:t}}\big[\big\|S_t^{(i)}\big\|_2^{2+\alpha}\big] < \infty$ *for all* $t \in [2\colon T]$.

Proof of Lemma C.1.1. Note that for any $\beta_{1:t-1} \in \mathbb{R}^{d_{1:t-1}}$,

$$\big|\pi_t\big(1, S_t^{(i)}; \beta_{1:t-1}\big) - \pi_t\big(1, S_t^{(i)}; \beta^*_{1:t-1}\big)\big|$$

$$= \left| \text{Clip}_{\pi_{\min}} \left[\text{expit}\left(\rho \cdot \beta_{t-1,1}^{\top} S_t^{(i)}\right) \right] - \text{Clip}_{\pi_{\min}} \left[\text{expit}\left(\rho \cdot (\beta_{t-1,1}^{*})^{\top} S_t^{(i)}\right) \right] \right|. \tag{C.1.5}$$

Note that for any real numbers x, y that $\left|\text{Clip}_{\pi_{\min}}(x) - \text{Clip}_{\pi_{\min}}(y)\right| \leq |x - y|$. This is because

- If $x, y \in [\pi_{\min}, 1 - \pi_{\min}]$, then $\left|\text{Clip}_{\pi_{\min}}(x) - \text{Clip}_{\pi_{\min}}(y)\right| = |x - y|$.
- If $x, y < \pi_{\min}$ or $x, y > 1 - \pi_{\min}$, then $0 = \left|\text{Clip}_{\pi_{\min}}(x) - \text{Clip}_{\pi_{\min}}(y)\right| \leq |x - y|$.
- If $x > \pi_{\min}$ and $y < \pi_{\min}$, then $\left|\text{Clip}_{\pi_{\min}}(x) - \text{Clip}_{\pi_{\min}}(y)\right| \leq x - \text{Clip}_{\pi_{\min}}(y) < x - y = |x - y|$. Same argument goes for the case that $y > \pi_{\min}$ and $x < \pi_{\min}$.
- If $x < 1 - \pi_{\min}$ and $y > 1 - \pi_{\min}$, then $\left|\text{Clip}_{\pi_{\min}}(x) - \text{Clip}_{\pi_{\min}}(y)\right| \leq \text{Clip}_{\pi_{\min}}(y) - x < y - x = |x - y|$. Same argument goes for the case that $y < 1 - \pi_{\min}$ and $x > 1 - \pi_{\min}$.

Thus,

$$\leq \left| \text{expit}\left(\rho \cdot \beta_{t-1,1}^{\top} S_t^{(i)}\right) - \text{expit}\left(\rho \cdot (\beta_{t-1,1}^{*})^{\top} S_t^{(i)}\right) \right| \tag{C.1.6}$$

Note that for any function $f : \mathcal{X} \mapsto [0, 1]$, for any $x, x' \in \mathcal{X}$,

$$|f(x) - f(x')| \leq \sup_{x_0 \in \mathcal{X}} \left| \frac{\partial}{\partial x} f(x) \Big|_{x=x_0} \right| |x - x'|.$$

Above $\sup_{x_0 \in \mathcal{X}} \left| \frac{\partial}{\partial x} f(x) \Big|_{x=x_0} \right|$ is the maximum absolute value of the derivative of f. We can use the above observation to upper bound display (C.1.6) as follows:

$$\leq \sup_{x_0 \in \mathbb{R}} \left| \frac{\partial}{\partial x} \text{expit}(x) \Big|_{x=x_0} \right| \cdot \left| \rho \cdot \beta_{t-1,1}^{\top} S_t^{(i)} - \rho \cdot (\beta_{t-1,1}^{*})^{\top} S_t^{(i)} \right| \tag{C.1.7}$$

Note that $\frac{\partial}{\partial x}\text{expit}(x) = \text{expit}(x)\{1 - \text{expit}(x)\}$ and that $\sup_{p\in[0,1]} p(1-p) = 0.25$. Thus,

$$\sup_{x_0\in\mathbb{R}}\left|\frac{\partial}{\partial x}\text{expit}(x)\Big|_{x=x_0}\right| = \sup_{x_0\in\mathbb{R}}\left|\text{expit}(x_0)\{1-\text{expit}(x_0)\}\right| \leq \sup_{p\in[0,1]}\left|p(1-p)\right| \leq 0.25.$$

Thus, we can upper bound display (C.1.7) with the following

$$\leq 0.25\left|\rho\cdot\beta_{t-1,1}^\top S_t^{(i)} - \rho\cdot(\beta_{t-1,1}^*)^\top S_t^{(i)}\right|$$

$$= 0.25\rho\left|\beta_{t-1,1}^\top S_t^{(i)} - (\beta_{t-1,1}^*)^\top S_t^{(i)}\right|$$

By Cauchy Schwartz inequality,

$$\leq 0.25\rho\left\|\beta_{t-1,1} - \beta_{t-1,1}^*\right\|_2\left\|S_t^{(i)}\right\|_2.$$

By the above result, Condition 5.3.3 holds since $\mathbb{E}\left[\left\|S_t^{(i)}\right\|_2^{2+\alpha}\right] < \infty$. ■

Stochastic Mirror Descent Algorithm (Lemma C.1.2)

We now give an example of an online stochastic mirror descent algorithm, based on those from[63] pg 361 and[16], whose policy class satisfies Conditions 5.3.2 and 5.3.3. We assume a binary action setting with $\mathcal{A} = \{0,1\}$ and assume that $\hat{\beta}_{t-1}^{(n)} = [\hat{\beta}_{t-1,0}^{(n)}, \hat{\beta}_{t-1,1}^{(n)}]$ below are estimated using the least squares criterion from display (5.3.8).

Note that for online stochastic mirror descent algorithms, $\hat{\pi}_t^{(n)}$ is an updated version of $\hat{\pi}_{t-1}^{(n)}$, which itself is an updated version of $\hat{\pi}_{t-2}^{(n)}$, and so on. This means that parameters of the class π_t must include those of $\pi_{t-1}, \pi_{t-2}, \ldots, \pi_2$. We will use slightly non-standard notation to represent this, $\hat{\pi}_t^{(n)}(\,\cdot\,) = \pi_t(\,\cdot\,;\hat{\beta}_{1:t-1}^{(n)})$, where each $\hat{\beta}_{t-1}^{(n)} = [\hat{\beta}_{t-1,0}^{(n)}, \hat{\beta}_{t-1,1}^{(n)}]$ is

estimated using the least squares criterion from display (5.3.8). Since we consider a binary action setting, to characterize a policy it is sufficient to define the probability that action 1 is selected in each state.

$$\begin{aligned}\hat{\pi}_t^{(n)}\big(1, S_t^{(i)}\big) &= \pi_t\big(1, S_t^{(i)}; \hat{\beta}_{1:t-1}^{(n)}\big) \\ &= \operatorname{argmin}_{p\in[\pi_{\min}, 1-\pi_{\min}]} \Big\{ \eta_t\big(-\hat{\beta}_{t-1,0}^{(n),\top} S_t^{(i)} - p\, \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)}\big) + \big(\hat{\pi}_{t-1}\big(1, S_t^{(i)}\big) - p\big)^2 \Big\}. \end{aligned} \tag{C.1.8}$$

Above, $\eta_t > 0$ is a learning rate and $\pi_{\min} \in (0, 0.5]$ is the minimum exploration rate. Note that $\hat{\beta}_{t-1,0}^{(n),\top} S_t^{(i)} + p\, \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)}$ is an estimate of the expectation of $R_t^{(i)}$ given $S_t^{(i)}, \mathcal{H}_{t-1}^{(i)}$ when $A_t^{(i)}$ is selected with probability p. Since the algorithm is designed to minimize a loss, we multiply $\hat{\beta}_{t-1,0}^{(n),\top} S_t^{(i)} + p\, \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)}$ by minus 1 to ensure the algorithm is maximizing the reward (i.e., minimizing the negative reward). The term $\big(\hat{\pi}_{t-1}(1, S_t^{(i)}) - p\big)^2$ is a Bregman divergence and can be replaced by other Bregman divergences, e.g., KL-divergence.

By display (C.1.8) above, we can derive

$$\pi_t\big(1, S_t^{(i)}; \hat{\beta}_{1:t-1}^{(n)}\big) = \operatorname{Clip}_{\pi_{\min}}\left(\pi_{t-1}\big(1, S_t^{(i)}; \hat{\beta}_{1:t-2}^{(n)}\big) + \frac{1}{2}\eta_t \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)} \right), \tag{C.1.9}$$

where $\operatorname{Clip}_{\pi_{\min}}(x) \triangleq \min\big(\max\big(x, \pi_{\min}\big), 1 - \pi_{\min}\big)$; see Lemma C.1.2 below for proof.

Note that exploration Condition 5.3.2 is satisfied because the action selection probabilities are constrained between $[\pi_{\min}, 1 - \pi_{\min}]$. We can also show that Condition 5.3.3 holds because

$$\big|\pi_t\big(1, S_t^{(i)}; \beta_{1:t-1}\big) - \pi_t\big(1, S_t^{(i)}; \beta_{1:t-1}^*\big)\big| \le \frac{1}{2}\eta_t \big\|S_t^{(i)}\big\| \, \big\|\beta_{t-1,1} - \beta_{t-1,1}^*\big\| \tag{C.1.10}$$

for any $\beta_{1:t-1} \in \mathbb{R}^{d_{1:t-1}}$, where $d_{1:t-1} \triangleq \sum_{t'=1}^{t-1} d_{t'}$; we also show this in Lemma C.1.2 below.

Lemma C.1.2 (Stochastic Mirror Descent Algorithm)**.** *We consider the stochastic mirror descent algorithm example that selects actions as described in display* (C.1.8)*. We show that display* (C.1.9) *holds. We also show that Condition 5.3.3 holds under the conditions that*

(a) $\mathbb{E}_{\pi^*_{2:t}}\left[\left\|S_t^{(i)}\right\|_2^{2+\alpha}\right] < \infty$ *for all* $t \in [2\colon T]$ *(the constant* $\alpha > 0$ *is the same as that from Condition 5.3.3), and*

(b) the learning rates are bounded, i.e., for a constant $\eta_{\max}$, $\eta_t \leq \eta_{\max} < \infty$ *for all* $t \in [1\colon T]$.

Proof of Lemma C.1.2.

Showing display (C.1.9) ***holds.*** Recall from display (C.1.8) that the stochastic mirror descent algorithm uses the following action selection probabilities:

$$\begin{aligned}\hat{\pi}_t^{(n)}\big(1, S_t^{(i)}\big) &= \pi_t\big(1, S_t^{(i)}; \hat{\beta}_{1:t-1}^{(n)}\big) \\ &= \operatorname{argmin}_{p \in [\pi_{\min}, 1-\pi_{\min}]} \left\{ \eta_t\big(-\hat{\beta}_{t-1,0}^{(n),\top} S_t^{(i)} - p\, \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)}\big) + \big(\hat{\pi}_{t-1}\big(1, S_t^{(i)}\big) - p\big)^2 \right\}.\end{aligned}$$

By taking the derivative of the following criterion with respect to p,

$$-\eta_t\big(\hat{\beta}_{t-1,0}^{(n),\top} S_t^{(i)} + p\, \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)}\big) + \big(\hat{\pi}_{t-1}\big(1, S_t^{(i)}\big) - p\big)^2, \tag{C.1.11}$$

we have

$$-\eta_t\, \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)} - 2\big\{\hat{\pi}_{t-1}\big(1, S_t^{(i)}\big) - p\big\}.$$

Since the second derivative of the criterion from display (C.1.11) with respect to p is $2 > 0$, the global minimizer of the criterion (not restricted to $[\pi_{\min}, 1 - \pi_{\min}]$) is $p =$

$\hat{\pi}_{t-1}\big(1, S_t^{(i)}\big) + \frac{1}{2}\eta_t \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)}$. Also note that the criterion from Equation (C.1.11) is convex because its derivative is strictly increasing in p. Note that the constrained minimizer of a convex function either equals the global minimizer or is on the boundary of the constraint space. Thus we have that the constrained minimizer, $\hat{\pi}_t\big(1, S_t^{(i)}\big)$, equals the following:

$$\pi_t\big(1, S_t^{(i)}; \hat{\beta}_{1:t-1}^{(n)}\big) = \text{Clip}_{\pi_{\min}}\left(\hat{\pi}_{t-1}(1, S_t^{(i)}) + \frac{1}{2}\eta_t \hat{\beta}_{t-1,1}^{(n),\top} S_t^{(i)}\right),$$

where $\text{Clip}_{\pi_{\min}}(x) \triangleq \min\big(\max(x, \pi_{\min}), 1 - \pi_{\min}\big)$. Thus, we have shown that display (C.1.9) holds.

Showing Condition 5.3.3 holds. Note that for any $\beta_{1:t-1} \in \mathbb{R}^{d_{1:t-1}}$,

$$\left|\pi_t\big(1, S_t^{(i)}; \beta_{1:t-1}\big) - \pi_t\big(1, S_t^{(i)}; \beta_{1:t-1}^*\big)\right|$$

$$= \left|\text{Clip}_{\pi_{\min}}\left(\hat{\pi}_{t-1}(1, S_t^{(i)}) + \frac{1}{2}\eta_t \beta_{t-1,1}^{\top} S_t^{(i)}\right) - \text{Clip}_{\pi_{\min}}\left(\hat{\pi}_{t-1}(1, S_t^{(i)}) + \frac{1}{2}\eta_t \beta_{t-1,1}^{*,\top} S_t^{(i)}\right)\right|.$$

Note that for any real numbers x, y that $\left|\text{Clip}_{\pi_{\min}}(x) - \text{Clip}_{\pi_{\min}}(y)\right| \leq |x - y|$; the justification for this is discussed below display (C.1.5). Thus, we have that display (C.1.5) can be upper bounded by the following:

$$\leq \left|\hat{\pi}_{t-1}(1, S_t^{(i)}) + \frac{1}{2}\eta_t \beta_{t-1,1}^{\top} S_t^{(i)} - \hat{\pi}_{t-1}(1, S_t^{(i)}) - \frac{1}{2}\eta_t \beta_{t-1,1}^{*,\top} S_t^{(i)}\right|$$

$$= \left|\frac{1}{2}\eta_t \left(\beta_{t-1,1} - \beta_{t-1,1}^*\right)^{\top} S_t^{(i)}\right| \leq \frac{1}{2}\eta_t \left\|S_t^{(i)}\right\|_2 \left\|\beta_{t-1,1} - \beta_{t-1,1}^*\right\|_2$$

The last inequality above holds by Cauchy-Schwartz. By the above, Condition 5.3.3 holds since $\mathbb{E}\left[\left\|S_t^{(i)}\right\|_2^{2+\alpha}\right] < \infty$ and $\eta_t \leq \eta_{\max} < \infty$ for all $t \in [1\colon T]$. ■

C.1.3 Radon-Nikodym Derivatives (Lemma C.1.3)

Lemma C.1.3 (Radon-Nikodym Derivatives)**.** *For any $\beta_{t-1} \in \mathbb{R}^{d_{t-1}}$, conditional on any $S_t^{(i)}$, $\mu(\,\cdot\,) \triangleq \pi_t(\,\cdot\,, S_t^{(i)}; \beta_{t-1})$ defines a probability measure on the sigma-algebra $\sigma(A_t^{(i)})$. Also conditional on any $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$, the policy $\nu(\,\cdot\,) \triangleq \hat{\pi}_t^{(n)}(\,\cdot\,, S_t^{(i)})$ defines a probability measure on the sigma-algebra $\sigma(A_t^{(i)})$.*

Under Condition 5.3.2, conditionally on almost every $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$, $g(\,\cdot\,) \triangleq \frac{\pi_t(\,\cdot\,, S_t^{(i)}; \beta_{t-1})}{\hat{\pi}_t^{(n)}(\,\cdot\,, S_t^{(i)})}$ is a Radon-Nikodym derivative, i.e., for any measurable subset $\bar{A} \subseteq \mathcal{A}$, conditionally on almost every $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$, $\mu(\bar{A}) = \int_{\bar{A}} g \, d\nu$.

Proof of Lemma C.1.3. We first show that that conditionally on almost every $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$, $\mu(\,\cdot\,) = \pi_t(\,\cdot\,, S_t^{(i)}; \beta_{t-1})$ is absolutely continuous with respect to $\nu(\,\cdot\,)$ $= \hat{\pi}_t^{(n)}(\,\cdot\,, S_t^{(i)})$.

By exploration Condition 5.3.2 we have that for any measurable subset measurable subset $\bar{A} \subseteq \mathcal{A}$, $\hat{\pi}_t^{(n)}(\bar{A}, S_t^{(i)}) \geq \pi_{\min} > 0$ a.s. This means that for any measurable subset $\bar{A} \subseteq \mathcal{A}$, $\hat{\pi}_t^{(n)}(\bar{A}, S_t^{(i)}) \geq \pi_{\min} > 0$ conditionally on almost every $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$. Thus we have that conditional on almost every $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$, $\mu(\,\cdot\,) = \pi_t(\,\cdot\,, S_t^{(i)}; \beta_{t-1})$ is absolutely continuous with respect to $\nu(\,\cdot\,) = \hat{\pi}_t^{(n)}(\,\cdot\,, S_t^{(i)})$.

Thus, for some function $g : \mathcal{A} \mapsto [0, \infty)$, conditionally on almost every $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$, $\mu(\bar{A}) = \int_{\bar{A}} g \, d\nu$ for any measurable subset $\bar{A} \subseteq \sigma(\mathcal{A})$. This means that conditionally on almost every $\mathcal{H}_{t-1}^{(1:n)}$ and $S_t^{(i)}$,

$$\pi_t(\bar{A}, S_t^{(i)}; \beta_{t-1}) = \int_{\bar{A}} g \, d\hat{\pi}_t^{(n)}(\,\cdot\,, S_t^{(i)}).$$

For $g(\,\cdot\,) = \frac{\pi_t(\,\cdot\,, S_t^{(i)}; \beta_{t-1})}{\hat{\pi}_t^{(n)}(\,\cdot\,, S_t^{(i)})}$, the above equality is satisfied. ■

C.1.4 Bracketing Numbers

Definition of Bracketing Numbers

Following the notation used in Chapter 19 of Van der Vaart [99], for any function class $\mathcal{F}$ of real-valued functions of $\mathcal{H}_T^{(i)}$, we use $N_{[\,]}\big(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)$ to denote the number of brackets of size ε in $L_P(\mathcal{P}_{\pi^*})$ norm needed to cover $\mathcal{F}$. Formally, this means we can find $N_\varepsilon \triangleq N_{[\,]}\big(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)$ number of brackets or pairs of real-valued functions of $\mathcal{H}_T^{(i)}$, $\big\{(l_k, u_k)\big\}_{k=1}^{N_\varepsilon}$, such that (i) all brackets together cover $\mathcal{F}$, i.e., for any $f \in \mathcal{F}$ we can find some bracket (l_k, u_k) such that $l_k(\mathcal{H}_T^{(i)}) \leq f(\mathcal{H}_T^{(i)}) \leq u_k(\mathcal{H}_T^{(i)})$ a.s., and (ii) the brackets have size less than ε, i.e., $\mathbb{E}_{\pi^*_{2:T}}\big[\big|u_k(\mathcal{H}_T^{(i)}) - l_k(\mathcal{H}_T^{(i)})\big|^p\big]^{1/p} < \varepsilon$. As done in Chapter 19 of Van der Vaart [99], we assume that the bracketing functions themselves also have finite $L_p(\mathcal{P}_{\pi^*})$ norm, i.e., for any $\varepsilon > 0$ and $k \in [1\colon N_\varepsilon]$, $\mathbb{E}_{\pi^*_{2:T}}\big[|u_k(\mathcal{H}_T^{(i)})|^p\big]^{1/p} < \infty$ and $\mathbb{E}_{\pi^*_{2:T}}\big[|l_k(\mathcal{H}_T^{(i)})|^p\big]^{1/p} < \infty$.

Product of Lipschitz Policy Functions are Lipschitz (Lemma C.1.4)

Lemma C.1.4 (Product of Lipschitz Policy Functions are Lipschitz)**.** *Let $t \in [3\colon T-1]$. Under Condition 5.3.3 (Lipschitz Policy Function), for any $\beta_{1:t-1}, \tilde{\beta}_{1:t-1} \in B_{1:t-1}$,*

$$\Bigg|\prod_{t'=2}^{t} \pi_{t'}\big(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}\big) - \prod_{t'=2}^{t} \pi_{t'}\big(A_{t'}^{(i)}, S_{t'}^{(i)}; \tilde{\beta}_{t'-1}\big)\Bigg| \leq \Bigg\{\sum_{t'=2}^{t} \dot{\pi}_{t'}\big(A_{t'}^{(i)}, S_{t'}^{(i)}\big)\Bigg\} \big\|\beta_{1:t-1} - \tilde{\beta}_{1:t-1}\big\|_2 \quad \text{a.s.} \quad \text{(C.1.12)}$$

Proof of Lemma C.1.4. Note that by telescoping series,

$$\prod_{t'=2}^{t} \pi_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}\right) - \prod_{t'=2}^{t} \pi_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}; \tilde{\beta}_{t'-1}\right)$$

$$= \left[\pi_2(A_2^{(i)}, S_2^{(i)}; \beta_1) - \pi_2(A_2^{(i)}, S_2^{(i)}; \tilde{\beta}_1)\right] \left\{ \prod_{t'=3}^{t} \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}) \right\}$$

$$+ \left\{\pi_2(A_2^{(i)}, S_2^{(i)}; \tilde{\beta}_1)\right\} \left[\pi_3(A_3^{(i)}, S_3^{(i)}; \beta_2) - \pi_3(A_3^{(i)}, S_3^{(i)}; \tilde{\beta}_2)\right] \left\{ \prod_{t'=4}^{t} \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}) \right\}$$

$$+ \dots$$

$$+ \left\{ \prod_{t'=2}^{t-1} \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \tilde{\beta}_{t'-1}) \right\} \left[\pi_t(A_t^{(i)}, S_t^{(i)}; \beta_{t-1}) - \pi_t(A_t^{(i)}, S_t^{(i)}; \tilde{\beta}_{t-1})\right]$$

$$= \sum_{t'=2}^{t} \left\{ \prod_{k=2}^{t'-1} \pi_k(A_k^{(i)}, S_k^{(i)}; \tilde{\beta}_{k-1}) \right\} \left[\pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}) - \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \tilde{\beta}_{t'-1})\right]$$

$$\left\{ \prod_{k=t'+1}^{t} \pi_k(A_k^{(i)}, S_k^{(i)}; \beta_{k-1}) \right\}$$

By slight abuse of notation, above we use $\prod_{k=2}^{1} \pi_k(A_k^{(i)}, S_k^{(i)}; \tilde{\beta}_{k-1}) = 1$ and $\prod_{k=t+1}^{t} \pi_k(A_k^{(i)}, S_k^{(i)}; \beta_{k-1}) = 1$.

Using the above result and triangle inequality,

$$\left| \prod_{t'=2}^{t} \pi_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}\right) - \prod_{t'=2}^{t} \pi_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}; \tilde{\beta}_{t'-1}\right) \right|$$

$$\leq \sum_{t'=2}^{t} \left| \prod_{k=2}^{t'-1} \pi_k(A_k^{(i)}, S_k^{(i)}; \tilde{\beta}_{k-1}) \right| \left| \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}) - \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \tilde{\beta}_{t'-1}) \right| \left| \prod_{k=t'+1}^{t} \pi_k(A_k^{(i)}, S_k^{(i)}; \beta_{k-1}) \right|$$

Since the terms $\left|\pi_k(A_k^{(i)}, S_k^{(i)}; \beta_{k-1})\right|$ and $\left|\pi_k(A_k^{(i)}, S_k^{(i)}; \tilde{\beta}_{k-1})\right|$ are less than or equal to 1 a.s.,

$$\leq \sum_{t'=2}^{t} \left| \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}) - \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \tilde{\beta}_{t'-1}) \right|$$

By Condition 5.3.3 (Lipschitz Policy Function),

$$\leq \sum_{t'=2}^{t} \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \left\| \beta_{t'-1} - \tilde{\beta}_{t'-1} \right\|_2$$

$$\leq \left\{ \sum_{t'=2}^{t} \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right\} \left\| \beta_{1:t-1} - \tilde{\beta}_{1:t-1} \right\|_2$$

By the above argument, we have that

$$\left| \prod_{t'=2}^{t} \pi_t(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}) - \prod_{t'=2}^{t} \pi_t(A_{t'}^{(i)}, S_{t'}^{(i)}; \tilde{\beta}_{t'-1}) \right| \leq \left\{ \sum_{t'=2}^{t} \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right\} \left\| \beta_{1:t-1} - \tilde{\beta}_{1:t-1} \right\|_2 \text{ a.s. } \blacksquare \quad \text{(C.1.13)}$$

Bracketing Number for Product of Function Classes (Lemma C.1.5)

Lemma C.1.5 (Bracketing Number for Product of Function Classes)**.** *Let $t \in [2 : T-1]$. Let $\mathcal{F}$ be a class of real-valued functions of $\mathcal{H}_t^{(i)}$ indexed by $\lambda \in L \subseteq \mathbb{R}^{d_L}$, i.e., $\mathcal{F} \triangleq \left\{ f(\cdot\,; \lambda) : \lambda \in L \right\}$. Also let $\Pi_{2:t} \triangleq \left\{ \prod_{t'=2}^{t} \pi_{t'}\left(\cdot\,; \beta_{t'-1}\right) \text{ s.t. } \beta_{1:t-1} \in B_{1:t-1} \right\}$. Finally, also*

let

$$\Pi_{2:t} \cdot \mathcal{F} \triangleq \left\{ \left[\prod_{t'=2}^{t} \pi_{t'}\left(\cdot\,; \beta_{t'-1}\right) \right] f(\cdot\,; \lambda) \text{ s.t. } \beta_{1:t-1} \in B_{1:t-1}, \lambda \in L \right\}.$$

For any $p \geq 1$, $N_{[\,]}\left(\varepsilon, \Pi_{2:t} \cdot \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\right) < \infty$ *for all* $\varepsilon > 0$ *under the following conditions:*

(i) Condition 5.3.3 (Lipschitz Policy Function) holds; if $p > 2$*, then Condition 5.3.3 must hold for* $\alpha = p - 2$*.*

(ii) $N_{[\,]}\left(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\right) < \infty$ *for all* $\varepsilon > 0$*.*

(iii) There exists a real-valued, measurable function F *of* $\mathcal{H}_t^{(i)}$ *such that (a)* $\left|f(\mathcal{H}_t^{(i)})\right| \leq F(\mathcal{H}_t^{(i)})$ *a.s. for all* $f \in \mathcal{F}$*, and (b)* $\mathbb{E}_{\pi^*_{2:t}}\left[\left|F(\mathcal{H}_t^{(i)})\dot{\pi}_{t'}(\mathcal{H}_{t'}^{(i)})\right|^p\right] < \infty$ *for all* $t' \in [2\colon t]$*, where the functions* $\dot{\pi}_{t'}$ *are from Condition 5.3.3.*

Furthermore, $\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon, \Pi_{2:t} \cdot \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\right)}d\varepsilon < \infty$ *under the additional condition that*

(iv) $\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\right)}d\varepsilon < \infty$

Note that assumption (iv) above implies that assumption (ii) holds.

Remark C.1.1 (Bracketing Number for the Product of Policy and Estimating Functions)**.** *Let* $\mathcal{F}_{\Pi c^\top \psi}\left(B_{1:T-1}, K_\theta\right) \triangleq \left\{ \left[\prod_{t=2}^{T} \pi_t(\cdot\,; \beta_{t-1})\right] c^\top \psi(\cdot\,; \theta) \text{ s.t. } \beta_{1:T-1} \in B_{1:T-1}, \theta \in K_\theta \right\}$ *for any compact set* $K_\theta \subseteq \mathbb{R}^{d_\theta}$ *and any* $c \in \mathbb{R}^{d_\theta}$*. By Lemma C.1.5, we have the following results:*

(a1) Under Condition 5.3.3 and assumption (C3), $\mathcal{F}_{\Pi c^\top \psi}\left(B_{1:T-1}, K_\theta\right)$ *has a finite bracketing number, i.e., that* $N_{[\,]}\left(\varepsilon,\ \mathcal{F}_{\Pi c^\top \psi}(B_{1:T-1}, K_\theta),\ L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ *for any* $\varepsilon > 0$*.*

(a2) Under Condition 5.3.3 and assumption (N4), $\mathcal{F}_{\Pi c^\top \psi}\big(B_{1:T-1}, \Theta\big)$ has a finite bracketing integral, i.e., that $\int_0^1 \sqrt{\log N_{[\,]}\Big(\varepsilon,\ \mathcal{F}_{\Pi c^\top \psi}(B_{1:T-1}, \Theta),\ L_{2+\alpha}(\mathcal{P}_{\pi^*})\Big)}d\varepsilon < \infty$ for any $\varepsilon > 0$.

Let $\mathcal{F}_{\Pi c^\top \varphi_t}(B_{1:t-1}, K_t) \triangleq \Big\{\big[\prod_{t'=2}^{t} \pi_{t'}(\,\cdot\,;\beta_{t'-1})\big] c^\top \varphi_t(\,\cdot\,;\beta_t)$ s.t. $\beta_{1:t-1} \in B_{1:t-1}, \beta_t \in K_t\Big\}$ for any compact set $K_t \subseteq \mathbb{R}^{d_t}$ and any $c \in \mathbb{R}^{d_t}$. By Lemma C.1.5, we have the following results:

(b1) Under Condition 5.3.3 and assumption (CP3), $\mathcal{F}_{\Pi c^\top \varphi_t}\big(B_{1:t-1}, K_t\big)$ has a finite bracketing number, i.e., that $N_{[\,]}\Big(\varepsilon,\ \mathcal{F}_{\Pi c^\top \varphi_t}(B_{1:t-1}, K_t),\ L_{1+\alpha}(\mathcal{P}_{\pi^*})\Big) < \infty$ for any $\varepsilon > 0$.

(b2) Under Condition 5.3.3 and assumption (NP1), $\mathcal{F}_{\Pi c^\top \varphi_t}\big(B_{1:t}\big)$ has a finite bracketing integral, i.e., that $\int_0^1 \sqrt{\log N_{[\,]}\Big(\varepsilon,\ \mathcal{F}_{\Pi c^\top \varphi_t}(B_{1:t}),\ L_{2+\alpha}(\mathcal{P}_{\pi^*})\Big)}d\varepsilon < \infty$ for any $\varepsilon > 0$.

Proof of Lemma C.1.5. For notational convenience, let

$$\pi_{2:t}(\mathcal{H}_t^{(i)}; \beta_{1:t-1}) \triangleq \prod_{t'=2}^{t} \pi_{t'}\big(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}\big).$$

By Lemma C.1.4 (Product of Lipschitz Policy Function are Lipschitz) for any $\beta_{1:t-1}, \tilde{\beta}_{1:t-1} \in B_{1:t-1}$,

$$\big|\pi_{2:t}(\mathcal{H}_t^{(i)}; \beta_{1:t-1}) - \pi_{2:t}(\mathcal{H}_t^{(i)}; \beta_{1:t-1})\big| \le \dot{\pi}_{2:t}(\mathcal{H}_t^{(i)})\big\|\beta_{1:t-1} - \tilde{\beta}_{1:t-1}\big\|_2 \tag{C.1.14}$$

where $\dot{\pi}_{2:t}(\mathcal{H}_t^{(i)}) \triangleq \sum_{t'=2}^{t} \dot{\pi}_{t'}\big(A_{t'}^{(i)}, S_{t'}^{(i)}\big)$. Note that by assumption (i) (that Condition 5.3.3 holds), $\|\dot{\pi}_{2:t}\|_{\mathcal{P}_{\pi^*}, p} \triangleq \mathbb{E}_{\pi_{2:t}^*}\big[|\dot{\pi}_{2:t}(\mathcal{H}_t^{(i)})|^p\big]^{1/p} < \infty$.

Constructing bracketing functions that cover $\Pi_{2:t}$**.** Let $\varepsilon > 0$. We now use the approach from Lemma 19.7 of[99]. Since $B_{1:t-1}$ is compact, the size of $B_{1:t-1}$ in every fixed dimension is at most $\operatorname{diam}(B_{1:t-1}) < \infty$. We can cover $B_{1:t-1}$ with $\left[\operatorname{diam}(B_{1:t-1})/\varepsilon\right]^{d_{1:t-1}}$ or fewer cubes of with edges of size ε; recall that $d_{1:t-1} \triangleq \sum_{t'=1}^{t-1} d_{t'}$. Let the projection of the centers of each of these finitely many cubes onto $B_{1:t-1}$ be the points $\mathcal{B}_{1:t-1} \subset B_{1:t-1}$. Note that $|\mathcal{B}_{1:t-1}| = \left[\operatorname{diam}(B_{1:t-1})/\varepsilon\right]^{d_{1:t-1}}$. For each of these cubes, consider the circumscribed ball that contains the cube; each of these balls has radius of $c\varepsilon$ for some constant $0 < c < \infty$.

We now construct a collection of bracketing functions that cover Π. These bracketing functions are

$$\left\{ \left[\pi_{2:t}\left(\,\cdot\,; \beta_{1:t-1}\right) - \varepsilon \dot{\pi}_{2:t}(\,\cdot\,),\ \pi_{2:t}\left(\,\cdot\,; \beta_{1:t-1}\right) + \varepsilon \dot{\pi}_{2:t}(\,\cdot\,) \right] \right\}_{\beta_{1:t-1} \in \mathcal{B}_{1:t-1}}. \tag{C.1.15}$$

The above brackets are of size at most $2\varepsilon \|\dot{\pi}_{2:t}\|_{\mathcal{P}_{\pi^*},p}$ in $L_p(\mathcal{P}_{\pi^*})$ norm by assumption (i) (that Condition 5.3.3 holds).

We now discuss why the brackets from display (C.1.15) cover $\Pi_{2:t}$. Consider any $\beta_{1:t-1} \in B_{1:t-1}$. Since the grid of cubes that cover $B_{1:t-1}$ whose projected centers form the collection of points $\mathcal{B}_{1:t-1}$ have edges of length ε, thus there must exists some $\beta^{(k)}_{1:t-1} \in \mathcal{B}_{1:t-1}$ such that $\left\|\beta_{1:t-1} - \beta^{(k)}_{1:t-1}\right\|_2 \leq \varepsilon$. By display (C.1.14),

$$\pi_{2:t}\left(\mathcal{H}_t^{(i)}; \beta_{1:t-1}\right) \in \left[\pi_{2:t}\left(\mathcal{H}_t^{(i)}; \beta^{(k)}_{1:t-1}\right) - \varepsilon \dot{\pi}_{2:t}(\mathcal{H}_t^{(i)}), \pi_{2:t}\left(\mathcal{H}_t^{(i)}; \beta^{(k)}_{1:t-1}\right) + \varepsilon \dot{\pi}_{2:t}(\mathcal{H}_t^{(i)}) \right] \text{ a.s.}$$

Thus, we have that

$$N_{[\,]}\left(2\varepsilon\|\dot{\pi}_{2:t}\|_{\mathcal{P}_{\pi^*},p}, \Pi_{2:t}\cdot\mathcal{F}, L_p(\mathcal{P}_{\pi^*})\right) \leq |\mathcal{B}_{1:t-1}| = \left[\operatorname{diam}(B_{1:t-1})/\varepsilon\right]^{d_{1:t-1}}. \tag{C.1.16}$$

Additionally, note that $0 \leq \pi_{2:t}(\mathcal{H}_t^{(i)};\beta_{1:t-1}) \leq 1$ a.s. since this function is a product of probabilities. Thus the brackets from display (C.1.15) can be modified such that the bracketing functions are in $[0,1]$ w.p. 1 while maintaining coverage of $\Pi_{2:t}$ and not increasing the size of the brackets. Specifically, these brackets are:

$$\begin{aligned}&\left\{\left(l_{\Pi,k}, u_{\Pi,k}\right)\right\}_{k=1}^{|\mathcal{B}_{1:t-1}|} \triangleq \\ &\left\{\left[\max\left\{0, \pi_{2:t}(\,\cdot\,;\beta_{1:t-1}) - \varepsilon\pi_{2:t}(\,\cdot\,)\right\}, \min\left\{1, \pi_{2:t}(\,\cdot\,;\beta_{1:t-1}) + \varepsilon\pi_{2:t}(\,\cdot\,)\right\}\right]\right\}_{\beta_{1:t-1}\in\mathcal{B}_{1:t-1}}.\end{aligned} \tag{C.1.17}$$

Constructing bracketing functions that cover $\Pi_{2:t}\cdot\mathcal{F}$**.** By assumption (ii), we can find $N_{\mathcal{F},\varepsilon} \triangleq N_{[\,]}\left(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\right) < \infty$ bracketing functions which cover $\mathcal{F}$. We will call these bracketing functions $\left\{\left[l_{\mathcal{F},k}, u_{\mathcal{F},k}\right]\right\}_{k=1}^{N_{\mathcal{F},\varepsilon}}$. We now show that we can construct a finite collection of bracketing functions which cover $\Pi_{2:t}\cdot\mathcal{F}$ using the bracketing functions for $\mathcal{F}$ and $\Pi_{2:t}$.

Consider any function $\pi_{2:t}(\,\cdot\,;\beta_{1:t-1})f(\,\cdot\,) \in \Pi_{2:t}\cdot\mathcal{F}$.

- From display (C.1.17), we can find some bracket $\left[l_{\Pi,k}, u_{\Pi,k}\right]$ for $k \in [1\colon |\mathcal{B}_{1:t-1}|]$ such that $0 \leq l_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq \pi_{2:t}(\mathcal{H}_t^{(i)};\beta_{1:t-1}) \leq u_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq 1$ a.s.
- Additionally, we can find some bracket $\left[l_{\mathcal{F},j}, u_{\mathcal{F},j}\right]$ for $k \in [1\colon N_{\mathcal{F},\varepsilon}]$ such that $l_{\mathcal{F},j}(\mathcal{H}_t^{(i)}) \leq f(\mathcal{H}_t^{(i)}) \leq u_{\mathcal{F},j}(\mathcal{H}_t^{(i)})$ a.s.

Now note the following observations for any particular $\mathcal{H}_t^{(i)}$:

- If for a particular $\mathcal{H}_t^{(i)}, f(\mathcal{H}_t^{(i)}) \geq 0$, then since $0 \leq l_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq u_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq 1$, then

$$l_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot l_{\mathcal{F},j}(\mathcal{H}_t^{(i)}) \leq \pi_{2:t}(\mathcal{H}_t^{(i)}; \beta_{1:t-1}) f(\mathcal{H}_t^{(i)}) \leq u_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot u_{\mathcal{F},j}(\mathcal{H}_t^{(i)}).$$

- If for a particular $\mathcal{H}_t^{(i)}, f(\mathcal{H}_t^{(i)}) < 0$, then since $0 \leq l_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq u_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq 1$, then

$$u_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot l_{\mathcal{F},j}(\mathcal{H}_t^{(i)}) \leq \pi_{2:t}(\mathcal{H}_t^{(i)}; \beta_{1:t-1}) f(\mathcal{H}_t^{(i)}) \leq l_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot u_{\mathcal{F},j}(\mathcal{H}_t^{(i)}).$$

By the above two observations, we have that

$$\begin{aligned} \min \left\{ l_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot l_{\mathcal{F},j}(\mathcal{H}_t^{(i)}), u_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot l_{\mathcal{F},j}(\mathcal{H}_t^{(i)}) \right\} &\leq \pi_{2:t}(\mathcal{H}_t^{(i)}; \beta_{1:t-1}) f(\mathcal{H}_t^{(i)}) \\ &\leq \max \left\{ u_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot u_{\mathcal{F},j}(\mathcal{H}_t^{(i)}), l_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot u_{\mathcal{F},j}(\mathcal{H}_t^{(i)}) \right\} \text{ a.s.} \end{aligned}$$

Thus, the following bracketing functions cover $\Pi_{2:t} \cdot \mathcal{F}$:

$$\begin{aligned} &\left\{ \left(l_{\Pi_{2:t} \cdot \mathcal{F},k}, u_{\Pi_{2:t} \cdot \mathcal{F},k} \right) \right\}_{k=1}^{|\mathcal{B}_{1:t-1}| \cdot N_{\mathcal{F},\varepsilon}} \\ &\quad \triangleq \left\{ \left[\min \left(l_{\Pi,k} \cdot l_{\mathcal{F},j}, u_{\Pi,k} \cdot l_{\mathcal{F},j} \right), \max \left(u_{\Pi,k} \cdot u_{\mathcal{F},j}, l_{\Pi,k} \cdot u_{\mathcal{F},j} \right) \right] \right\}_{k=1;j=1}^{k=|\mathcal{B}_{1:t-1}|;j=N_{\mathcal{F},\varepsilon}}. \end{aligned} \tag{C.1.18}$$

Note that there are $|\mathcal{B}_{1:t-1}| \cdot N_{\mathcal{F},\varepsilon}$ brackets above.

We now derive the size of the above brackets for $\Pi_{2:t} \cdot \mathcal{F}$.

$$\left|\max\left(u_{\Pi,k} \cdot u_{\mathcal{F},j}, l_{\Pi,k} \cdot u_{\mathcal{F},j}\right) - \min\left(l_{\Pi,k} \cdot l_{\mathcal{F},j}, u_{\Pi,k} \cdot l_{\mathcal{F},j}\right)\right|$$

Since $\left|\max(a,b) - c\right| \leq |a-c| + |b-c|$,

$$\leq \left|u_{\Pi,k} \cdot u_{\mathcal{F},j} - \min\left(l_{\Pi,k} \cdot l_{\mathcal{F},j}, u_{\Pi,k} \cdot l_{\mathcal{F},j}\right)\right| + \left|l_{\Pi,k} \cdot u_{\mathcal{F},j} - \min\left(l_{\Pi,k} \cdot l_{\mathcal{F},j}, u_{\Pi,k} \cdot l_{\mathcal{F},j}\right)\right| \text{ a.s.}$$

$$\leq \left|u_{\Pi,k} \cdot u_{\mathcal{F},j} - l_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + \left|u_{\Pi,k} \cdot u_{\mathcal{F},j} - u_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + \left|l_{\Pi,k} \cdot u_{\mathcal{F},j} - l_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + \left|l_{\Pi,k} \cdot u_{\mathcal{F},j} - u_{\Pi,k} \cdot l_{\mathcal{F},j}\right| \text{ a.s.}$$

Since $0 \leq u_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq 1$ a.s. and $0 \leq l_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq 1$ a.s.,

$$= \left|u_{\Pi,k} \cdot u_{\mathcal{F},j} - l_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + \left|l_{\Pi,k} \cdot u_{\mathcal{F},j} - u_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + u_{\Pi,k}\left|u_{\mathcal{F},j} - l_{\mathcal{F},j}\right| + l_{\Pi,k}\left|u_{\mathcal{F},j} - l_{\mathcal{F},j}\right|$$

Again since $0 \leq u_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq 1$ a.s. and $0 \leq l_{\Pi,k}(\mathcal{H}_t^{(i)}) \leq 1$,

$$\leq \left|u_{\Pi,k} \cdot u_{\mathcal{F},j} - l_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + \left|l_{\Pi,k} \cdot u_{\mathcal{F},j} - u_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + 2\left|u_{\mathcal{F},j} - l_{\mathcal{F},j}\right| \text{ a.s.}$$

By triangle inequality,

$$\leq \left|u_{\Pi,k} \cdot u_{\mathcal{F},j} - u_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + \left|u_{\Pi,k} \cdot l_{\mathcal{F},j} - l_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + \left|l_{\Pi,k} \cdot u_{\mathcal{F},j} - l_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + \left|l_{\Pi,k} \cdot l_{\mathcal{F},j} - u_{\Pi,k} \cdot l_{\mathcal{F},j}\right| + 2\left|u_{\mathcal{F},j} - l_{\mathcal{F},j}\right| \text{ a.s.}$$

Using the same arguments as used above,

$$\leq \left|u_{\mathcal{F},j} - l_{\mathcal{F},j}\right| + |u_{\Pi,k} - l_{\Pi,k}|\,|l_{\mathcal{F},j}| + \left|u_{\mathcal{F},j} - l_{\mathcal{F},j}\right| + |l_{\Pi,k} - u_{\Pi,k}|\,|l_{\mathcal{F},j}| + 2\left|u_{\mathcal{F},j} - l_{\mathcal{F},j}\right| \text{ a.s.}$$

$$= 2\,|u_{\Pi,k} - l_{\Pi,k}|\,|l_{\mathcal{F},j}| + 4\left|u_{\mathcal{F},j} - l_{\mathcal{F},j}\right|$$

Thus,

$$\mathbb{E}_{\pi^*_{2:t}}\Bigg[\Bigg|\max\left\{u_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot u_{\mathcal{F},j}(\mathcal{H}_t^{(i)}), l_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot u_{\mathcal{F},j}(\mathcal{H}_t^{(i)})\right\} - \min\left\{l_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot l_{\mathcal{F},j}(\mathcal{H}_t^{(i)}), u_{\Pi,k}(\mathcal{H}_t^{(i)}) \cdot l_{\mathcal{F},j}(\mathcal{H}_t^{(i)})\right\}\Bigg|^p\Bigg]$$

$$\leq \mathbb{E}_{\pi^*_{2:t}}\left[\left|2\left|u_{\Pi,k}(\mathcal{H}_t^{(i)}) - l_{\Pi,k}(\mathcal{H}_t^{(i)})\right|\left|l_{\mathcal{F},j}(\mathcal{H}_t^{(i)})\right| + 4\left|u_{\mathcal{F},j}(\mathcal{H}_t^{(i)}) - l_{\mathcal{F},j}(\mathcal{H}_t^{(i)})\right|\right|^p\right]$$

By Lemma C.2.1 (Inequality using Binomial Theorem), for some constant $c_p < \infty$,

$$\leq 2^p c_p \mathbb{E}_{\pi^*_{2:t}}\left[\left|u_{\Pi,k}(\mathcal{H}_t^{(i)}) - l_{\Pi,k}(\mathcal{H}_t^{(i)})\right|^p |l_{\mathcal{F},j}(\mathcal{H}_t^{(i)})|^p\right] + 4^p c_p \mathbb{E}_{\pi^*_{2:t}}\left[\left|u_{\mathcal{F},j}(\mathcal{H}_t^{(i)}) - l_{\mathcal{F},j}(\mathcal{H}_t^{(i)})\right|^p\right]$$

Since brackets $[l_{\mathcal{F},j}, u_{\mathcal{F},j}]$ are of size ε or less in $L_p(\mathcal{P}_{\pi^*})$ norm by construction,

$$\leq 2^p c_p \mathbb{E}_{\pi^*_{2:t}} \left[\left| u_{\Pi,k}(\mathcal{H}_t^{(i)}) - l_{\Pi,k}(\mathcal{H}_t^{(i)}) \right|^p \left| l_{\mathcal{F},j}(\mathcal{H}_t^{(i)}) \right|^p \right] + 4^p c_p \varepsilon^p$$

By the definition of brackets $[l_{\Pi,k}, u_{\Pi,k}]$ from display (C.1.17),

$$\leq 2^p c_p \mathbb{E}_{\pi^*_{2:t}} \left[\left| 2\varepsilon \dot{\pi}_{2:t}(\mathcal{H}_t^{(i)}) \right|^p |l_{\mathcal{F},j}(\mathcal{H}_t^{(i)})|^p \right] + 4^p c_p \varepsilon^p$$

Recall that by assumption of the Lemma, F is a function such that $|f(\mathcal{H}_t^{(i)})| \leq F(\mathcal{H}_t^{(i)})$ a.s. for all $f \in \mathcal{F}$. Thus, the brackets $\left[l_{\mathcal{F},k}, u_{\mathcal{F},k}\right]$ can always be chosen such that $|l_{\mathcal{F},j}(\mathcal{H}_t^{(i)})| \leq F(\mathcal{H}_t^{(i)})$ a.s. Thus,

$$\leq 2^p c_p \mathbb{E}_{\pi^*_{2:t}} \left[\left| 2\varepsilon \dot{\pi}_{2:t}(\mathcal{H}_t^{(i)}) \right|^p |F(\mathcal{H}_t^{(i)})|^p \right] + 4^p c_p \varepsilon^p$$

Since $\dot{\pi}_{2:t}(\mathcal{H}_t^{(i)}) \geq 0$ and $F(\mathcal{H}_t^{(i)}) \geq 0$ by definition,

$$= 2^{2p} c_p \varepsilon^p \mathbb{E}_{\pi^*_{2:t}} \left[\left| \dot{\pi}_{2:t}(\mathcal{H}_t^{(i)}) F(\mathcal{H}_t^{(i)}) \right|^p \right] + 4^p c_p \varepsilon^p$$

Since $\dot{\pi}_{2:t}(\mathcal{H}_t^{(i)}) = \sum_{t'=2}^{t} \dot{\pi}_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}\right)$ by definition,

$$= 2^{2p} c_p \varepsilon^p \mathbb{E}_{\pi^*_{2:t}} \left[\left| \sum_{t'=2}^{t} \dot{\pi}_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}\right) F(\mathcal{H}_t^{(i)}) \right|^p \right] + 4^p c_p \varepsilon^p$$

By repeatedly applying Lemma C.2.1 (Inequality using Binomial Theorem), for some posi-

tive constant k_p,

$$= \varepsilon^p \left\{ 2^{2p} c_p k_p^t \sum_{t'=2}^{t} \mathbb{E}_{\pi^*_{2:t}} \left[\left| \dot{\pi}_{t'} \left(A_{t'}^{(i)}, S_{t'}^{(i)} \right) F(\mathcal{H}_t^{(i)}) \right|^p \right] + 4^p c_p \right\}.$$

The term $2^{2p} c_p k_p^t \sum_{t'=2}^{t} \mathbb{E}_{\pi^*_{2:t}} \left[\left| \dot{\pi}_{t'} \left(A_{t'}^{(i)}, S_{t'}^{(i)} \right) F(\mathcal{H}_t^{(i)}) \right|^p \right] + 4^p c_p$ above is bounded by assumption (iii).

Let $c_{\Pi \cdot \mathcal{F}} \triangleq \left\{ 2^{2p} c_p k_p^t \sum_{t'=2}^{t} \mathbb{E}_{\pi^*_{2:t}} \left[\left| \dot{\pi}_{t'} \left(A_{t'}^{(i)}, S_{t'}^{(i)} \right) F(\mathcal{H}_t^{(i)}) \right|^p \right] + 4^p c_p \right\}^{1/p}$. By the above result, the brackets for $\Pi_{2:t} \cdot \mathcal{F}$ from display (C.1.18) have size at most $\varepsilon c_{\Pi \cdot \mathcal{F}}$ in $L_p(\mathcal{P}_{\pi^*})$ norm. Thus,

$$\begin{aligned} N_{[\,]} \left(\varepsilon c_{\Pi \cdot \mathcal{F}}, \Pi_{2:t} \cdot \mathcal{F}, L_p(\mathcal{P}_{\pi^*}) \right) \leq & \underbrace{|\mathcal{B}_{1:t-1}|}_{\text{Upper bounds } N_{[\,]}\left(\varepsilon, \Pi_{2:t}, L_p(\mathcal{P}_{\pi^*})\right)} \cdot N_{\mathcal{F},\varepsilon} \\ & = \left[\operatorname{diam}(B_{1:t-1})/\varepsilon \right]^{d_{1:t-1}} N_{[\,]} \left(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*}) \right). \end{aligned}$$

The final equality above holds by display (C.1.16).

The above implies that

$$N_{[\,]} \left(\varepsilon, \Pi_{2:t} \cdot \mathcal{F}, L_p(\mathcal{P}_{\pi^*}) \right) \leq \left[\operatorname{diam}(B_{1:t-1}) c_{\Pi \cdot \mathcal{F}} / \varepsilon \right]^{d_{1:t-1}} \cdot N_{[\,]} \left(\varepsilon / c_{\Pi \cdot \mathcal{F}}, \mathcal{F}, L_p(\mathcal{P}_{\pi^*}) \right). \tag{C.1.19}$$

Note that by assumption (ii), $N_{[\,]} \left(\varepsilon / c_{\Pi \cdot \mathcal{F}}, \mathcal{F}, L_p(\mathcal{P}_{\pi^*}) \right) < \infty$, so the above implies that $N_{[\,]} \left(\varepsilon, \Pi_{2:t} \cdot \mathcal{F}, L_p(\mathcal{P}_{\pi^*}) \right) < \infty$ for all $\varepsilon > 0$.

Bracketing integral result. We now work on showing the second part of the Lemma,

i.e., that $\int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon, \Pi_{2:t} \cdot \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)}d\varepsilon < \infty$.

$$\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon, \Pi_{2:t} \cdot \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\right)}d\varepsilon$$

By display (C.1.19),

$$\leq \int_0^1 \sqrt{\log \left\{ \left[\text{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon\right]^{d_{1:t-1}} N_{[\,]}\big(\varepsilon/c_{\Pi\cdot\mathcal{F}}, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)\right\}}d\varepsilon$$

Using properties of log,

$$= \int_0^1 \sqrt{d_{1:t-1}\log\left[\text{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon\right] + \log N_{[\,]}\big(\varepsilon/c_{\Pi\cdot\mathcal{F}}, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)}d\varepsilon$$

Note that $\text{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon \geq 1$ since $\text{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon$ upper bounds the bracketing number $N_{[\,]}\left(\varepsilon, \Pi_{2:t}, L_p(\mathcal{P}_{\pi^*})\right)$ (which must be at least 1). Thus, $\log\left[\text{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon\right]$ ≥ 0. Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any numbers $a, b > 0$ (to see this, square both sides of the inequality),

$$= \sqrt{d_{1:t-1}}\int_0^1 \sqrt{\log\left[\text{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon\right]}d\varepsilon + \int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon/c_{\Pi\cdot\mathcal{F}}, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)}d\varepsilon. \tag{C.1.20}$$

We now discuss why the quantity in the display above is finite:

- Regarding the first term in display (C.1.20), note that since $\log\left[\text{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon\right]$

≥ 0, we have that $\log\left[\operatorname{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon\right] \leq \operatorname{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon$. Thus,

$$\int_0^1 \sqrt{\log\left[\operatorname{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon\right]}d\varepsilon \leq \int_0^1 \sqrt{\operatorname{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}/\varepsilon}d\varepsilon$$
$$= \sqrt{\operatorname{diam}(B_{1:t-1})c_{\Pi\cdot\mathcal{F}}}\int_0^1 \varepsilon^{-2}d\varepsilon < \infty.$$

- For the second term in display (C.1.20) above, note that

$$\int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon/c_{\Pi\cdot\mathcal{F}}, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)}d\varepsilon = c_{\Pi\cdot\mathcal{F}}\int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon/c_{\Pi\cdot\mathcal{F}}, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)}c_{\Pi\cdot\mathcal{F}}^{-1}d\varepsilon$$

  By integration by substitution, for $u = \varepsilon/c_{\Pi\cdot\mathcal{F}}$,

$$= c_{\Pi\cdot\mathcal{F}}\int_0^{c_{\Pi\cdot\mathcal{F}}^{-1}} \sqrt{\log N_{[\,]}\big(u, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)}du < \infty. \tag{C.1.21}$$

  The above term is finite since $\int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)}d\varepsilon < \infty$ by assumption (iv). If $c_{\Pi\cdot\mathcal{F}}^{-1} \leq 1$ display (C.1.21) holds straightforwardly. If $c_{\Pi\cdot\mathcal{F}}^{-1} \geq 1$, display (C.1.21) holds because $N_{[\,]}\big(1, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big) \geq N_{[\,]}\big(u, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)$ for all $u \geq 1$ by the definition of bracketing numbers.

We have shown that $\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon, \Pi_{2:t}\cdot\mathcal{F}, L_p(\mathcal{P}_{\pi^*})\right)}d\varepsilon < \infty$, so we have now shown the second result of the Lemma. ■

C.2 Policy Parameter Results

Overview of Supplement C.2 Results.

- **Section C.2.1:** Consistency of Policy Parameters (Sufficient Assumptions for Condition 5.3.1; Theorem C.2.1)
- **Section C.2.2:** Inequality Using Binomial Theorem (Helper Lemma C.2.1)
- **Section C.2.3:** Stochastic Equicontinuity for Policy Parameters (Theorem C.2.2)
- **Section C.2.4:** Invertibility of $\dot{\Phi}^*_{1:t}$ (Lemma C.2.2)

C.2.1 Consistency of Policy Parameters (Theorem C.2.1)

Theorem C.2.1 (Consistency of Policy Parameters)**.** *We assume Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Function) hold. Condition 5.3.1 holds (i.e., $\hat{\beta}_t^{(n)} \xrightarrow{P} \beta_t^*$ for each $t \in [1\colon T-1]$), under the following additional assumptions:*

(CP1) ***Well-Separated Solutions:*** *For each $t \in [1\colon T-1]$, for any $\varepsilon > 0$, there exists some $\eta > 0$ such that*

$$\inf_{\beta_t \in \mathbb{R}^{d_t} \text{ s.t. } \|\beta_t - \beta_t^*\|_1 > \varepsilon} \left\| \mathbb{E}_{\pi^*_{2:t}}\left[\varphi_t(\mathcal{H}_t^{(i)}; \beta_t)\right] \right\|_1 > \eta > 0.$$

(CP2) ***Asymptotically Tight:*** *For each $t \in [1\colon T-1]$, for any $\varepsilon > 0$, there exists some $k < \infty$ such that*

$$\limsup_{n \to \infty} \mathbb{P}\left(\left\| \hat{\beta}_t^{(n)} \right\|_1 > k \right) \leq \varepsilon.$$

(CP3) ***Finite Bracketing Number:*** *Let $\alpha > 0$ be a constant. For each $t \in [1\colon T-1]$ and any compact subset $K_t \subset \mathbb{R}^{d_t}$,*

(i) For any $\varepsilon > 0$ and any vector $c \in \mathbb{R}^{d_t}$, the bracketing number

$$N_{[\,]}\left(\varepsilon,\ \left\{c^\top \varphi_t(\,\cdot\,;\beta_t)\right\}_{\beta_t \in K_t},\ L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty.$$

(ii) There exists a function F_{φ_t} such that for all $\beta_t \in K_t$, $\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right\|_1 \le F_{\varphi_t}(\mathcal{H}_t^{(i)})$ a.s. and

$$\mathbb{E}_{\pi^*_{2:t}}\left[\left|F_{\varphi_t}(\mathcal{H}_t^{(i)})\dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)})\right|^{1+\alpha}\right] < \infty$$

for all $t' \in [2\colon t]$; the functions $\dot{\pi}_{t'}$ are from Condition 5.3.3.

Proof of Theorem C.2.1 (Consistency of Policy Parameters). We use an induction-based argument. For the base case, we show that $\hat{\beta}_1^{(n)} \xrightarrow{P} \beta_1^*$. For the induction step, we show that $\hat{\beta}_t^{(n)} \xrightarrow{P} \beta_t^*$, given that $\hat{\beta}_{1:t-1}^{(n)} \xrightarrow{P} \beta_{1:t-1}^*$.

Base Case. Note that $\mathcal{H}_1^{(1)}, \mathcal{H}_1^{(2)}, \mathcal{H}_1^{(3)}, \ldots, \mathcal{H}_1^{(n)}$ are i.i.d. (data from the first time-step; no adaptive sampling yet). This type of consistency proof is standard for Z-estimators; we include it for completeness since the induction step uses similar techniques.

We first define the following useful functions:

$$\Phi_1(\beta_1) \triangleq \mathbb{E}\left[\varphi_1(\mathcal{H}_1^{(i)};\beta_1)\right] \quad \text{and} \quad \hat{\Phi}_1^{(n)}(\beta_1) \triangleq \frac{1}{n}\sum_{i=1}^{n} \varphi_1(\mathcal{H}_1^{(i)};\beta_1).$$

Let $\varepsilon > 0$. By assumption (CP1), there exists some $\eta > 0$ such that if $\beta_1 \in \mathbb{R}^{d_1}$ satisfies $\left\|\beta_1 - \beta_1^*\right\|_1 > \varepsilon$, then $\left\|\Phi_1(\beta_1)\right\|_1 = \left\|\mathbb{E}\left[\varphi_1(\mathcal{H}_1^{(i)};\beta_1)\right]\right\|_1 > \eta > 0$. Thus, $\mathbb{I}_{\|\hat{\beta}_1^{(n)} - \beta_1^*\|_1 > \varepsilon} \le$

$\mathbb{I}_{\|\Phi_1(\hat{\beta}1)\|_1>\eta}$, so

$$\mathbb{P}\left(\left\|\hat{\beta}_1^{(n)}-\beta_1^*\right\|_1>\varepsilon\right)\leq\mathbb{P}\left(\left\|\Phi_1\left(\hat{\beta}1\right)\right\|_1>\eta\right).$$

By the definition of $\hat{\beta}1$ from display (5.3.9), $\hat{\Phi}_1^{(n)}\left(\hat{\beta}_1^{(n)}\right) = \frac{1}{n}\sum_{i=1}^n \varphi_1\left(\mathcal{H}_1^{(i)};\hat{\beta}1\right) = o_P(1/\sqrt{n}) = o_P(1)$. Thus,

$$=\mathbb{P}\left(\left\|\hat{\Phi}_1^{(n)}\left(\hat{\beta}1\right)-\Phi_1\left(\hat{\beta}1\right)\right\|_1>\eta-o_P(1)\right).$$

By assumption (CP2), for any $\delta>0$, there exists some $k<\infty$ such that $\limsup_{n\to\infty}\mathbb{P}\left(\left\|\hat{\beta}_1^{(n)}\right\|_1>k\right)\leq\delta$. We use this k below:

$$=\mathbb{P}\left(\left\|\hat{\Phi}_1^{(n)}\left(\hat{\beta}_1^{(n)}\right)-\Phi_1\left(\hat{\beta}_1^{(n)}\right)\right\|_1\left\{\mathbb{I}_{\|\hat{\beta}_1^{(n)}\|_1>k}+\mathbb{I}_{\|\hat{\beta}_1^{(n)}\|_1\leq k}\right\}>\eta-o_P(1)\right)$$

$$\leq\mathbb{P}\left(\left\|\hat{\Phi}_1^{(n)}\left(\hat{\beta}_1^{(n)}\right)-\Phi_1\left(\hat{\beta}_1^{(n)}\right)\right\|_1\mathbb{I}_{\|\hat{\beta}_1^{(n)}\|_1\leq k}>\eta/2-o_P(1)\right)$$
$$+\mathbb{P}\left(\left\|\hat{\Phi}_1^{(n)}\left(\hat{\beta}_1^{(n)}\right)-\Phi_1\left(\hat{\beta}_1^{(n)}\right)\right\|_1\mathbb{I}_{\|\hat{\beta}_1^{(n)}\|_1>k}>\eta/2-o_P(1)\right)$$

$$\leq\mathbb{P}\left(\left\|\hat{\Phi}_1^{(n)}\left(\hat{\beta}_1^{(n)}\right)-\Phi_1\left(\hat{\beta}_1^{(n)}\right)\right\|_1\mathbb{I}_{\|\hat{\beta}_1^{(n)}\|_1\leq k}>\eta/2-o_P(1)\right)+\mathbb{P}\left(\left\|\hat{\beta}_1^{(n)}\right\|_1>k\right)+o(1)$$

$$\leq\underbrace{\mathbb{P}\left(\sup_{\beta_1\in\mathbb{R}^{d_1}\text{ s.t. }\|\beta_1\|_1\leq k}\left\|\hat{\Phi}_1^{(n)}\left(\beta_1\right)-\Phi_1\left(\beta_1\right)\right\|_1>\eta/2-o_P(1)\right)}_{=o(1)}+\underbrace{\mathbb{P}\left(\left\|\hat{\beta}_1^{(n)}\right\|_1>k\right)}_{\leq\delta}+o(1). \tag{C.2.1}$$

For the second term above, by assumption (CP2), $\limsup_{n\to\infty}\mathbb{P}\left(\left\|\hat{\beta}_1^{(n)}\right\|_1>k\right)\leq\delta$ and

δ can be made arbitrarily small.

For the first term above, we can apply the Uniform Weak Law of Large Numbers result for i.i.d. data [100] Theorem 2.4.1 to get that it converges to zero as $n \to \infty$. Specifically, note that

$$N_{[\,]}\left(\varepsilon,\ \left\{c^\top \varphi_1(\,\cdot\,;\beta_1)\right\}_{\beta_1 \in \mathbb{R}^{d_1} \text{ s.t. } \|\beta_1\|_1 \leq k},\ L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty,$$

for any $c \in \mathbb{R}^{d_1}$ and any $\varepsilon > 0$ by assumption (CP3). Thus, by the Uniform Weak Law of Large Numbers, we have that

$$\sup_{\beta_1 \in \mathbb{R}^{d_1} \text{ s.t. } \|\beta_1\|_1 \leq k} \left| c^\top \hat{\Phi}_1^{(n)}(\beta_1) - c^\top \Phi_1(\beta_1) \right| \xrightarrow{P} 0,$$

for any $c \in \mathbb{R}^{d_1}$. Thus, by Cramer Wold device we have that

$$\sup_{\beta_1 \in \mathbb{R}^{d_1} \text{ s.t. } \|\beta_1\|_1 \leq k} \left\| \hat{\Phi}_1^{(n)}(\beta_1) - \Phi_1(\beta_1) \right\|_1 \xrightarrow{P} 0.$$

Thus, the expression in display (C.2.1) above converges to zero.

Induction Step. For our induction assumption, we assume that $\hat{\beta}_{1:t-1}^{(n)} \xrightarrow{P} \beta_{1:t-1}^*$. Given this assumption, we will show that $\hat{\beta}_t^{(n)} \xrightarrow{P} \beta_t^*$.

We now define the following useful functions:

$$\Phi_t(\beta_{1:t}) \triangleq \mathbb{E}_{\pi(\beta_{1:t-1})}\left[\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right] = \mathbb{E}\left[W_{2:t}^{(i)}(\beta_{1:t-1},\hat{\beta}_{1:t-1}^{(n)})\,\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right] \tag{C.2.2}$$

and

$$\hat{\Phi}_t^{(n)}(\beta_{1:t}) \triangleq \frac{1}{n}\sum_{i=1}^{n} W_{2:t}^{(i)}(\beta_{1:t-1},\hat{\beta}_{1:t-1}^{(n)})\,\varphi_t(\mathcal{H}_t^{(i)};\beta_t). \tag{C.2.3}$$

For now, we take as given that display (C.2.4) below holds; we will show this result holds at the end of this proof.

$$\left\| \Phi_t\left(\hat{\beta}^{(n)}_{1:t-1}, \beta^*_t\right) - \Phi_t\left(\beta^*_{1:t-1}, \beta^*_t\right) \right\|_1 = o_P(1). \tag{C.2.4}$$

Let $\varepsilon > 0$. By assumption (CP1), there exists some $\eta > 0$ such that if $\beta_t \in \mathbb{R}^{d_t}$ satisfies $\left\|\beta_t - \beta^*_t\right\|_1 > \varepsilon$, then $\left\|\Phi_t\left(\beta^*_{1:t-1}, \beta_t\right)\right\|_1 = \left\|\mathbb{E}_{\pi^*_{2:t}}\left[\varphi_t(\mathcal{H}^{(i)}_t; \beta_t)\right]\right\|_1 > \eta > 0$. Thus,

$$\mathbb{P}\left(\left\|\hat{\beta}^{(n)}_t - \beta^*_t\right\|_1 > \varepsilon\right) \leq \mathbb{P}\left(\left\|\Phi_t\left(\beta^*_{1:t-1}, \hat{\beta}^{(n)}_t\right)\right\|_1 > \eta\right)$$

Note $\left\|\Phi_t\left(\beta^*_{1:t-1}, \hat{\beta}^{(n)}_t\right)\right\|_1 = \left\|\Phi_t\left(\beta^*_{1:t-1}, \hat{\beta}^{(n)}_t\right) - \Phi_t\left(\hat{\beta}^{(n)}_{1:t-1}, \hat{\beta}^{(n)}_t\right) + \Phi_t\left(\hat{\beta}^{(n)}_{1:t-1}, \hat{\beta}^{(n)}_t\right)\right\|_1 \leq \left\|\Phi_t\left(\beta^*_{1:t-1}, \hat{\beta}^{(n)}_t\right) - \Phi_t\left(\hat{\beta}^{(n)}_{1:t-1}, \hat{\beta}^{(n)}_t\right)\right\|_1 + \left\|\Phi_t\left(\hat{\beta}^{(n)}_{1:t-1}, \hat{\beta}^{(n)}_t\right)\right\|_1 = \left\|\Phi_t\left(\hat{\beta}^{(n)}_{1:t}\right)\right\|_1 + o_P(1)$; the last equality holds by display (C.2.4) and the definition of $\hat{\beta}^{(n)}_t$ from display (5.3.9). Thus,

$$\leq \mathbb{P}\left(\left\|\Phi_t\left(\hat{\beta}^{(n)}_{1:t}\right)\right\|_1 > \eta - o_P(1)\right)$$

Note $\hat{\Phi}^{(n)}_t\left(\hat{\beta}^{(n)}_{1:t}\right) = \frac{1}{n}\sum_{i=1}^n W^{(i)}_{2:t}\left(\hat{\beta}^{(n)}_{1:t-1}, \hat{\beta}^{(n)}_{1:t-1}\right)\varphi_t\left(\mathcal{H}^{(i)}_t; \hat{\beta}^{(n)}_t\right) = \frac{1}{n}\sum_{i=1}^n \varphi_t\left(\mathcal{H}^{(i)}_t; \hat{\beta}^{(n)}_t\right) = o_P(1/\sqrt{n}) = o_P(1)$; the second to last equality holds by the definition of $\hat{\beta}^{(n)}_t$ from display (5.3.9). Thus,

$$= \mathbb{P}\left(\left\|\hat{\Phi}^{(n)}_t\left(\hat{\beta}^{(n)}_{1:t}\right) - \Phi_t\left(\hat{\beta}^{(n)}_{1:t}\right)\right\|_1 > \eta - o_P(1)\right)$$

By assumption (CP2), for any $\delta > 0$, there exists some $k < \infty$ such that $\limsup_{n\to\infty} \mathbb{P}\left(\left\|\hat{\beta}^{(n)}_t\right\|_1 > k\right) \leq \delta$. We use this k below:

$$= \mathbb{P}\left(\left\|\hat{\Phi}^{(n)}_t\left(\hat{\beta}^{(n)}_{1:t}\right) - \Phi_t\left(\hat{\beta}^{(n)}_{1:t}\right)\right\|_1 \left\{\mathbb{I}_{\|\hat{\beta}^{(n)}_t\|_1 > k} + \mathbb{I}_{\|\hat{\beta}^{(n)}_t\|_1 \leq k}\right\} > \eta - o_P(1)\right)$$

$$\leq \mathbb{P}\left(\left\|\hat{\Phi}_t^{(n)}\left(\hat{\beta}_{1:t}^{(n)}\right) - \Phi_t\left(\hat{\beta}_{1:t}^{(n)}\right)\right\|_1 \mathbb{I}_{\|\hat{\beta}_t^{(n)}\|_1 \leq k} > \eta/2 - o_P(1)\right)$$
$$+ \mathbb{P}\left(\left\|\hat{\Phi}_t^{(n)}\left(\hat{\beta}_{1:t}^{(n)}\right) - \Phi_t\left(\hat{\beta}_{1:t}^{(n)}\right)\right\|_1 \mathbb{I}_{\|\hat{\beta}_t^{(n)}\|_1 > k} > \eta/2 - o_P(1)\right)$$

$$\leq \mathbb{P}\left(\left\|\hat{\Phi}_t^{(n)}\left(\hat{\beta}_{1:t}^{(n)}\right) - \Phi_t\left(\hat{\beta}_{1:t}^{(n)}\right)\right\|_1 \mathbb{I}_{\|\hat{\beta}_t^{(n)}\|_1 \leq k} > \eta/2 - o_P(1)\right) + \mathbb{P}\left(\left\|\hat{\beta}_t^{(n)}\right\|_1 > k\right) + o(1)$$

$$\leq \mathbb{P}\left(\sup_{\beta_t \in \mathbb{R}^{d_t} \text{ s.t. } \|\beta_t\|_1 \leq k} \left\|\hat{\Phi}_t^{(n)}\left(\hat{\beta}_{1:t-1}^{(n)}, \beta_t\right) - \Phi_t\left(\hat{\beta}_{1:t-1}^{(n)}, \beta_t\right)\right\|_1 > \eta/2 - o_P(1)\right)$$
$$+ \mathbb{P}\left(\left\|\hat{\beta}_t^{(n)}\right\|_1 > k\right) + o(1) \quad \text{(C.2.5)}$$

Recall $\hat{\beta}_{1:t-1}^{(n)} \xrightarrow{P} \beta_{1:t-1}^*$ by our induction assumption; thus, $\mathbb{I}_{\hat{\beta}_{1:t-1}^{(n)} \in B_{1:t-1}} \xrightarrow{P} 1$, where recall that $B_{1:t-1} \subset \mathbb{R}^{d_{1:t-1}}$ is a compact subset whose interior contains $\beta_{1:t-1}^*$. Thus,

$$\leq \underbrace{\mathbb{P}\left(\sup_{\beta_t \in \mathbb{R}^{d_t} \text{ s.t. } \|\beta_t\|_1 \leq k} \sup_{\beta_{1:t-1} \in B_{1:t-1}} \left\|\hat{\Phi}_t^{(n)}\left(\beta_{1:t-1}, \beta_t\right) - \Phi_t\left(\beta_{1:t-1}, \beta_t\right)\right\|_1 > \eta/2 - o_P(1)\right)}_{=o(1)}$$
$$+ \underbrace{\mathbb{P}\left(\left\|\hat{\beta}_t^{(n)}\right\|_1 > k\right)}_{\leq \delta} + o(1)$$

Note that the above converges to zero as $n \to \infty$ for the following reasons:

- By assumption (CP2), $\limsup_{n\to\infty} \mathbb{P}\left(\left\|\hat{\beta}_t^{(n)}\right\|_1 > k\right) \leq \delta$ and δ can be made arbitrarily small.

- Note that by Cramer Wold device, to show that

$$\sup_{\beta_t \in \mathbb{R}^{d_t} \text{ s.t. } \|\beta_t\|_1 \le k} \sup_{\beta_{1:t-1} \in B_{1:t-1}} \left\| \hat{\Phi}_t^{(n)}(\beta_{1:t-1}, \beta_t) - \Phi_t(\beta_{1:t-1}, \beta_t) \right\|_1 \xrightarrow{P} 0,$$

it is sufficient to show that for any vector $c \in \mathbb{R}^{d_t}$,

$$\sup_{\beta_t \in \mathbb{R}^{d_t} \text{ s.t. } \|\beta_t\|_1 \le k} \sup_{\beta_{1:t-1} \in B_{1:t-1}} c^\top \left\{ \hat{\Phi}_t^{(n)}(\beta_{1:t-1}, \beta_t) - \Phi_t(\beta_{1:t-1}, \beta_t) \right\} \xrightarrow{P} 0.$$

Also, note that

$$\sup_{\beta_t \in \mathbb{R}^{d_t} \text{ s.t. } \|\beta_t\|_1 \le k} \sup_{\beta_{1:t-1} \in B_{1:t-1}} c^\top \left\{ \hat{\Phi}_t^{(n)}(\beta_{1:t-1}, \beta_t) - \Phi_t(\beta_{1:t-1}, \beta_t) \right\}$$

$$= \sup_{\beta_t \in \mathbb{R}^{d_t} \text{ s.t. } \|\beta_t\|_1 \le k} \sup_{\beta_{1:t-1} \in B_{1:t-1}} \frac{1}{n} \sum_{i=1}^{n} \left\{ W_{2:t}^{(i)}(\beta_{1:t-1}, \hat{\beta}_{1:t-1}^{(n)}) c^\top \varphi_t(\mathcal{H}_t^{(i)}; \beta_t) \right.$$
$$\left. - \mathbb{E}\left[W_{2:t}^{(i)}(\beta_{1:t-1}, \hat{\beta}_{1:t-1}^{(n)}) c^\top \varphi_t(\mathcal{H}_t^{(i)}; \beta_t) \right] \right\} \xrightarrow{P} 0.$$

The above convergence result holds by Theorem C.4.2 (Weighted Martingale Triangular Array Uniform Weak Law of Large Numbers). Specifically we are able to apply Theorem C.4.2 because Condition 5.3.2 holds and because $N_{[\,]}\left(\varepsilon, \mathcal{F}_{\Pi c^\top \varphi_t}(B_{1:t-1}, k), L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ for all $c \in \mathbb{R}^{d_t}$ and all $\varepsilon > 0$, where

$$\mathcal{F}_{\Pi c^\top \varphi_t}(B_{1:t-1}, k) \triangleq \left\{ \left[\prod_{t'=2}^{t} \pi_{t'}(\,\cdot\,; \beta_{t'-1}) \right] c^\top \varphi_t(\,\cdot\,; \beta_t) \right\}_{\beta_{1:t-1} \in B_{1:t-1},\, \beta_t \in \mathbb{R}^{d_t} \text{ s.t. } \|\beta_t\|_1 \le k}.$$

The above finite bracketing number result holds since using assumption (CP3) and Condition 5.3.3, we can apply by Lemma C.1.5 (specifically see Remark C.1.1 part (b1)).

We now show that display (C.2.4) ***holds.*** Let $\beta_{1:t-1} \in B_{1:t-1}$.

$$\left\| \Phi_t\left(\beta_{1:t-1}, \beta_t^*\right) - \Phi_t\left(\beta_{1:t-1}^*, \beta_t^*\right) \right\|_1$$

$$= \left\| \mathbb{E}\left[\left\{ W_{2:t}^{(i)}\left(\beta_{1:t-1}, \hat{\beta}_{1:t-1}^{(n)}\right) - W_{2:t}^{(i)}\left(\beta_{1:t-1}^*, \hat{\beta}_{1:t-1}^{(n)}\right) \right\} \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right] \right\|_1$$

$$= \left\| \mathbb{E}_{\pi_{2:t}^*}\left[\left\{ W_{2:t}^{(i)}\left(\beta_{1:t-1}, \beta_{1:t-1}^*\right) - W_{2:t}^{(i)}\left(\beta_{1:t-1}^*, \beta_{1:t-1}^*\right) \right\} \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right] \right\|_1$$

By Jensen's inequality,

$$\leq \mathbb{E}_{\pi_{2:t}^*}\left[\left| W_{2:t}^{(i)}\left(\beta_{1:t-1}, \beta_{1:t-1}^*\right) - W_{2:t}^{(i)}\left(\beta_{1:t-1}^*, \beta_{1:t-1}^*\right) \right| \left\| \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right\|_1 \right]$$

By definition of $W_{t'}^{(i)}\left(\beta_{t'-1}, \beta_{t'-1}^*\right)$ from display (5.5.5),

$$= \mathbb{E}_{\pi_{2:t}^*}\left[\left| \prod_{t'=2}^{t} \pi_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}\right) - \prod_{t'=2}^{t} \pi_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}^*\right) \right| \right.$$
$$\left. \left\{ \prod_{t'=2}^{t} \frac{1}{\pi_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}^*\right)} \right\} \left\| \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right\|_1 \right]$$

By Condition 5.3.2 (Minimum Exploration), $\pi_{t'}\left(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}^*\right)^{-1} \leq \pi_{\min}^{-1}$ a.s. Thus,

$$\leq \pi_{\min}^{-(t-1)} \mathbb{E}_{\pi_{2:t}^*}\left[\left| \prod_{t'=2}^{t} \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}) - \prod_{t'=2}^{t} \pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}^*) \right| \left\| \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right\|_1 \right] \tag{C.2.6}$$

By Condition 5.3.3, we can apply Lemma C.1.4 (Product of Lipschitz Policy Functions are Lipschitz) to get that

$$\leq \pi_{\min}^{-(t-1)} \mathbb{E}_{\pi^*_{2:t}} \left[\left\| \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right\|_1 \left\{ \sum_{t'=2}^{t} \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right\} \left\| \beta_{1:t-1} - \beta^*_{1:t-1} \right\|_2 \right]$$

By linearity of expectations,

$$= \pi_{\min}^{-(t-1)} \left\{ \sum_{t'=2}^{t} \mathbb{E}_{\pi^*_{2:t}} \left[\left\| \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right\|_1 \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right] \right\} \left\| \beta_{1:t-1} - \beta^*_{1:t-1} \right\|_2.$$

Thus, by consolidating the above results, we have that

$$\begin{aligned} &\left\| \Phi_t\left(\hat{\beta}^{(n)}_{1:t-1}, \beta_t^*\right) - \Phi_t\left(\beta^*_{1:t-1}, \beta_t^*\right) \right\|_1 \\ &\quad \leq \pi_{\min}^{-(t-1)} \left\{ \sum_{t'=2}^{t} \mathbb{E}_{\pi^*_{2:t}} \left[\left\| \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right\|_1 \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right] \right\} \left\| \hat{\beta}^{(n)}_{1:t-1} - \beta^*_{1:t-1} \right\|_2 = o_P(1). \end{aligned}$$

The last limit above holds because

- $\left\| \hat{\beta}^{(n)}_{1:t-1} - \beta^*_{1:t-1} \right\|_2 = o_P(1)$ since $\hat{\beta}^{(n)}_{1:t-1} \xrightarrow{P} \beta^*_{1:t-1}$ by our induction assumption.
- By assumption (CP3) (Finite Bracketing Number for Policy Functions), there exists a function F_{φ_t} such that $\left\| \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t\right) \right\|_1 \leq F_{\varphi_t}(\mathcal{H}_t^{(i)})$ a.s. and for all $t' \in [2\colon t]$, $\mathbb{E}_{\pi^*_{2:t}}\left[F_{\varphi_t}\left(\mathcal{H}_t^{(i)}\right) \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right] < \infty$. Thus,

$$\mathbb{E}_{\pi^*_{2:t}} \left[\left\| \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right) \right\|_1 \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right] \leq \mathbb{E}_{\pi^*_{2:t}} \left[F_{\varphi_t}\left(\mathcal{H}_t^{(i)}\right) \dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right] < \infty.$$

We have now shown that display (C.2.4) holds. ■

C.2.2 Inequality Using Binomial Theorem (Helper Lemma C.2.1)

Lemma C.2.1 (Inequality Using Binomial Theorem)**.** *For any $\eta \geq 1$ and any $a, b \in \mathbb{R}$, we have that $|a+b|^\eta \leq c_\eta(|a|^\eta + |b|^\eta)$ for some constant $c_\eta < \infty$.*

Proof of Lemma C.2.1. Note that

$$|a+b|^\eta \leq \begin{cases} |a-b|^{\lfloor \eta \rfloor} & \text{if } |a+b| < 1 \\ |a-b|^{\lceil \eta \rceil} & \text{if } |a+b| \geq 1 \end{cases}$$

Above we use $\lfloor \eta \rfloor$ to round η down to the nearest integer and we use $\lceil \eta \rceil$ to round η up to the nearest integer. Let

$$k \triangleq \begin{cases} \lfloor \eta \rfloor & \text{if } |a+b| < 1 \\ \lceil \eta \rceil & \text{if } |a+b| \geq 1 \end{cases}.$$

Note $k \geq 1$ since that $\lceil \eta \rceil \geq \lfloor \eta \rfloor \geq 1$ because $\eta \geq 1$. Since k is a positive integer, by the Binomial theorem,

$$|a+b|^k = \left|(a+b)^k\right| = \left|\sum_{j=0}^{k} \binom{k}{j} a^j\, b^{k-j}\right| = \sum_{j=0}^{k} \binom{k}{j} |a|^j\, |b|^{k-j}$$

Note that $|a|^j\, |b|^{k-j} \leq \max\left(|a|^k, |b|^k\right)$ for all $j \in [0\colon k]$. Thus,

$$\leq \left\{\sum_{j=0}^{k} \binom{k}{j}\right\} \max\left(|a|^k, |b|^k\right) \leq \left\{\sum_{j=0}^{k} \binom{k}{j}\right\} \left(|a|^k + |b|^k\right).$$

Thus we can choose $c_\eta = \left\{\sum_{j=0}^{k} \binom{k}{j}\right\}$. ■

C.2.3 Asymptotic Equicontinuity for Policy Parameters (Theorem C.2.2)

Theorem C.2.2 (Asymptotic Equicontinuity for Policy Parameters)**.** *We assume that Conditions 5.3.1-5.5.1 on the adaptive sampling algorithm hold. Consider the following results:*

$$\sqrt{n}\left\{\hat{\Phi}^{(n)}_{1:T-1}(\hat{\beta}^{(n)}_{1:T-1}) - \Phi_{1:T-1}(\hat{\beta}^{(n)}_{1:T-1})\right\} = \sqrt{n}\left\{\hat{\Phi}^{(n)}_{1:T-1}(\beta^{*}_{1:T-1}) - \Phi_{1:T-1}(\beta^{*}_{1:T-1})\right\} + o_P(1) \quad \text{(C.2.7)}$$

and

$$\sqrt{n}(\hat{\beta}^{(n)}_{1:T-1} - \beta^{*}_{1:T-1}) = O_P(1).$$

The above results hold under the following additional assumptions:

(NP1) ***Finite Bracketing Integral:*** *Let $\alpha > 0$ be a constant. For each $t \in [1\colon T-1]$, for any vector $c \in \mathbb{R}^{d_t}$,*

$$\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon,\ \left\{c^\top \varphi_t(\,\cdot\,;\beta_t)\right\}_{\beta_t \in B_t},\ L_{2+\alpha}(\mathcal{P}_{\pi^*})\right)}\, d\varepsilon < \infty. \quad \text{(C.2.8)}$$

Additionally, there exists a function F_{φ_t} such that for all $\beta_t \in B_t$, $\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right\|_1 \le F_{\varphi_t}(\mathcal{H}_t^{(i)})$ a.s. and

$$\mathbb{E}_{\pi^*_{2:T}}\left[\left|F_{\varphi,t}(\mathcal{H}_t^{(i)})\dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)})\right|^{2+\alpha}\right] < \infty \quad \text{(C.2.9)}$$

for all $t' \in [2\colon T]$; the functions $\dot{\pi}_{t'}$ above are from Condition 5.3.3.

(NP2) ***Continuity Condition:*** *For each $t \in [1\colon T-1]$, the following mapping is continuous*

at $\beta_t = \beta_t^*$*:*

$$\beta_t \mapsto \mathbb{E}_{\pi^*_{2:t}}\left[\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t) - \varphi_t(\mathcal{H}_t^{(i)};\beta_t^*)\right\|_2^2\right].$$

Remark C.2.1 (Condition 5.5.2 implies assumptions (NP1) and (NP2) above hold)**.** *As we discussed in the Remark below Condition 5.5.2 (Lipschitz Policy Estimating Function), Condition 5.5.2 can be replaced by more general assumptions; these are assumptions (NP1) and (NP2) above.*

- *Condition 5.5.2 implies assumption (NP1) because*
  - *Example 19.7 of*[99] *shows that Lipschitz property of Condition 5.5.2 and the compactness of* B_t *implies that the bracketing integral condition from display* (C.2.8) *holds.*
  - *We now show why Condition 5.5.2 implies that display* (C.2.9) *holds. Since* B_t *is compact, let* $\text{diam}(B_t) < \infty$ *be the diameter of* B_t *in Euclidian distance. Note the following inequality for all* $\beta_t \in B_t$*,*

$$\begin{aligned}\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right\|_1 &\leq \sqrt{d_t}\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t)\right\|_2 \\ &\leq \sqrt{d_t}\left\{\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t^*)\right\|_2 + \dot{\varphi}_t(\mathcal{H}_t^{(i)})\text{diam}(B_t)\right\} \quad \text{a.s.}\end{aligned}$$

The first inequality above holds by property of norms and the second inequality above holds by display (5.5.18) *of Condition 5.3.3. Furthermore, note that*

$$\mathbb{E}_{\pi^*_{2:T}}\left[\left|\sqrt{d_t}\left\{\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t^*)\right\|_2 + \dot{\varphi}_t(\mathcal{H}_t^{(i)})\text{diam}(B_t)\right\}\dot{\pi}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)})\right|^{2+\alpha}\right].$$

By Lemma C.2.1 (Inequality Using Binomial Theorem), for some positive constant $c_{2+\alpha} < \infty$,

$$\leq c_{2+\alpha}\mathbb{E}_{\pi^*_{2:T}}\left[\left|\sqrt{d_t}\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t^*)\right\|_2\dot{\pi}_{t'}(A_{t'}^{(i)},S_{t'}^{(i)})\right|^{2+\alpha}\right]$$
$$+ c_{2+\alpha}\mathbb{E}_{\pi^*_{2:T}}\left[\left|\sqrt{d_t}\dot{\varphi}_t(\mathcal{H}_t^{(i)})\mathrm{diam}(B_t)\dot{\pi}_{t'}(A_{t'}^{(i)},S_{t'}^{(i)})\right|^{2+\alpha}\right] < \infty.$$

The final inequality holds by Condition 5.5.2. The above implies that display (C.2.9) *holds.*

- *We now discuss why Condition 5.5.2 implies assumption (NP2) holds. Let* $\varepsilon > 0$. *We want to show that there exists some* $\delta > 0$ *such that*

$$\mathbb{E}_{\pi^*_{2:t}}\left[\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t) - \varphi_t(\mathcal{H}_t^{(i)};\beta_t^*)\right\|_2^2\right] \leq \varepsilon$$

whenever $\left\|\beta_t - \beta_t^*\right\|_2 \leq \delta$. *By Condition 5.5.2, for any* $\beta_t \in B_t$,

$$\mathbb{E}_{\pi^*_{2:t}}\left[\left\|\varphi_t(\mathcal{H}_t^{(i)};\beta_t) - \varphi_t(\mathcal{H}_t^{(i)};\beta_t^*)\right\|_2^2\right] \leq \mathbb{E}_{\pi^*_{2:t}}\left[\dot{\varphi}_t(\mathcal{H}_t^{(i)})^2\right]\left\|\beta_t - \beta_t^*\right\|_2^2.$$

Recall $\mathbb{E}_{\pi^*_{2:t}}\left[\dot{\varphi}_t(\mathcal{H}_t^{(i)})^2\right] < \infty$ *by Condition 5.5.2. Thus,* $\delta = \varepsilon^{1/2}\mathbb{E}_{\pi^*_{2:t}}\left[\dot{\varphi}_t(\mathcal{H}_t^{(i)})^2\right]^{-1/2}$ *is sufficient.*

Proof of Theorem C.2.2. Recall in displays (5.5.8) and (5.5.9), we defined the functions $\Phi_{1:T-1}(\beta_{1:T-1})$ and $\hat{\Phi}^{(n)}_{1:T-1}(\beta_{1:T-1})$ respectively. More generally, we now define for *any*

$t \in [1\colon T-1]$ the following functions:

$$\Phi_{1:t}(\beta_{1:t}) \triangleq \mathbb{E}\begin{bmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1) \\ W_2^{(i)}(\beta_1,\hat{\beta}_1^{(n)})\varphi_2(\mathcal{H}_2^{(i)};\beta_2) \\ W_{2:3}^{(i)}(\beta_{1:2},\hat{\beta}_{1:2}^{(n)})\varphi_3(\mathcal{H}_3^{(i)};\beta_3) \\ \vdots \\ W_{2:t}^{(i)}(\beta_{1:t-1},\hat{\beta}_{1:t-1}^{(n)})\varphi_t(\mathcal{H}_t^{(i)};\beta_t) \end{bmatrix}$$

and

$$\hat{\Phi}_{1:t}^{(n)}(\beta_{1:t}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1) \\ W_2^{(i)}(\beta_1,\hat{\beta}_1^{(n)})\varphi_2(\mathcal{H}_2^{(i)};\beta_2) \\ W_{2:3}^{(i)}(\beta_{1:2},\hat{\beta}_{1:2}^{(n)})\varphi_3(\mathcal{H}_3^{(i)};\beta_3) \\ \vdots \\ W_{2:t}^{(i)}(\beta_{1:t-1},\hat{\beta}_{1:t-1}^{(n)})\varphi_t(\mathcal{H}_t^{(i)};\beta_t) \end{bmatrix}.$$

We use an induction argument.

- For the base case, $t = 1$, we will show that $\sqrt{n}\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) = O_P(1)$ and that

$$\sqrt{n}\left\{\hat{\Phi}_1^{(n)}(\beta 1) - \Phi_1(\beta 1)\right\} = \sqrt{n}\left\{\hat{\Phi}_1^{(n)}(\beta_1^*) - \Phi_1(\beta_1^*)\right\} + o_P(1).$$

- For the induction step, $t > 1$, we assume that $\sqrt{n}\big(\hat{\beta}_{1:t-1}^{(n)} - \beta_{1:t-1}^*\big) = O_P(1)$ and will show that $\sqrt{n}\big(\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big) = O_P(1)$ and that

$$\sqrt{n}\left\{\hat{\Phi}_{1:t}^{(n)}(\hat{\beta}_{1:t}^{(n)}) - \Phi_{1:t}(\hat{\beta}_{1:t}^{(n)})\right\} = \sqrt{n}\left\{\hat{\Phi}_{1:t}^{(n)}(\beta_{1:t}^*) - \Phi_{1:t}(\beta_{1:t}^*)\right\} + o_P(1).$$

To show that the Theorem holds, it is sufficient to show the above results hold.

Base Case. Note that since $\mathcal{H}_1^{(1)}, \mathcal{H}_1^{(2)}, \mathcal{H}_1^{(3)}, \ldots, \mathcal{H}_1^{(n)}$ are i.i.d., we can use an asymptotic normality argument for Z-estimators on i.i.d. data.

Stochastic Equicontinuity Result. First, we will apply Lemma C.6.1 (Stochastic Equicontinuity) to get that for any fixed vector $c \in \mathbb{R}^{d_1}$,

$$\sqrt{n}c^\top \left\{ \hat{\Phi}_1^{(n)}(\hat{\beta}_1^{(n)}) - \Phi_1(\hat{\beta}_1^{(n)}) \right\} = \sqrt{n}c^\top \left\{ \hat{\Phi}_1^{(n)}(\beta_1^*) - \Phi_1(\beta_1^*) \right\} + o_P(1). \tag{C.2.10}$$

We are able to apply Lemma C.6.1 (Stochastic Equicontinuity) because the following assumptions hold:

- Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) hold.
- By assumption (NP1) (Finite Bracketing Integral), for any vector $c \in \mathbb{R}^{d_1}$,

$$\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon,\ \left\{c^\top \varphi_1(\,\cdot\,;\beta_1)\right\}_{\beta_1 \in B_1},\ L_{2+\alpha}(\mathcal{P}_{\pi^*})\right)} d\varepsilon < \infty.$$

- $\hat{\beta}_1^{(n)} \xrightarrow{P} \beta_1^*$ by Condition 5.3.1. Thus, by assumption (NP2) and continuous mapping theorem we have that for any vector $c \in \mathbb{R}^{d_1}$, $\nu\big(c^\top \varphi_t(\,\cdot\,;\beta_t),\ c^\top \varphi_t(\,\cdot\,;\beta_t^*)\big) \xrightarrow{P} 0$, where

$$\nu\Big(c^\top \varphi_t(\,\cdot\,;\beta_t),\ c^\top \varphi_t(\,\cdot\,;\beta_t')\Big) \triangleq \mathbb{E}_{\pi^*_{2:t}}\left[\left\{c^\top \varphi_t\big(\mathcal{H}_t^{(i)};\beta_t\big) - c^\top \varphi_t\big(\mathcal{H}_t^{(i)};\beta_t'\big)\right\}^2\right]^{1/2}.$$

By display (C.2.10) and Cramer Wold device, we have that

$$\sqrt{n}\left\{\hat{\Phi}_1^{(n)}(\hat{\beta}_1^{(n)}) - \Phi_1(\hat{\beta}_1^{(n)})\right\} = \sqrt{n}\left\{\hat{\Phi}_1^{(n)}(\beta_1^*) - \Phi_1(\beta_1^*)\right\} + o_P(1). \tag{C.2.11}$$

With the above result, for the base case of the induction argument, we now just need to show that $\sqrt{n}\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) = O_P(1)$.

Showing that $\sqrt{n}\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) = O_P(1)$. By the definitions of $\hat{\beta}_1^{(n)}$ and β_1^* from displays (5.3.9) and (5.3.7) respectively, we can rewrite the left-hand side of the display (C.2.11) above:

$$\sqrt{n}\big\{\underbrace{\hat{\Phi}_1^{(n)}(\hat{\beta}_1^{(n)})}_{=o_P(1/\sqrt{n})} - \Phi_1(\hat{\beta}_1^{(n)})\big\} = \sqrt{n}\big\{\underbrace{\Phi_1(\beta_1^*)}_{=0} - \Phi_1(\hat{\beta}_1^{(n)})\big\} + o_P(1).$$

By Condition 5.5.1 (Differentiability of Policy Parameter Estimating Functions), the mapping $\beta_1 \mapsto \Phi_1(\beta_1)$ is differentiable at $\beta_1 = \beta_1^*$ with derivative matrix $\dot{\Phi}_1^* \triangleq \frac{\partial}{\partial \beta_1}\Phi_1(\beta_1)\big|_{\beta_1=\beta_1^*}$. So,

$$= -\sqrt{n}\dot{\Phi}_1^*\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) + \sqrt{n}o_P\big(\big\|\hat{\beta}_1^{(n)} - \beta_1^*\big\|_2\big) + o_P(1).$$

In summary, we have that

$$\sqrt{n}\left\{\hat{\Phi}_1^{(n)}(\beta_1^*) - \Phi_1(\beta_1^*)\right\} + o_P(1) = -\sqrt{n}\dot{\Phi}_1^*\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) + \sqrt{n}o_P\big(\big\|\hat{\beta}_1^{(n)} - \beta_1^*\big\|_2\big).$$

By Condition 5.5.1 (Differentiability of Policy Parameter Estimating Functions), $\dot{\Phi}_1^*$ is

invertible, so $\big[\dot{\Phi}_1^*\big]^{-1} = O(1)$ and

$$\sqrt{n}\big[\dot{\Phi}_1^*\big]^{-1}\Big\{\hat{\Phi}_1^{(n)}(\beta_1^*) - \Phi_1(\beta_1^*)\Big\} + o_P(1)$$
$$= -\sqrt{n}\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) + \sqrt{n}O(1)o_P\big(\big\|\hat{\beta}_1^{(n)} - \beta_1^*\big\|_2\big). \quad \text{(C.2.12)}$$

By the central limit theorem (for i.i.d. data),

$$\sqrt{n}\big[\dot{\Phi}_1^*\big]^{-1}\big\{\hat{\Phi}_1^{(n)}(\beta_1^*) - \underbrace{\Phi_1(\beta_1^*)}_{=0}\big\} = \frac{1}{\sqrt{n}}\big[\dot{\Phi}_1^*\big]^{-1}\sum_{i=1}^{n}\varphi_1\big(\mathcal{H}_1^{(i)};\beta_1^*\big) + o_P(1)$$
$$\xrightarrow{D} \mathcal{N}\left(0,\ \big[\dot{\Phi}_1^*\big]^{-1}\Sigma_1\big[\dot{\Phi}_1^*\big]^{-1,\top}\right), \quad \text{(C.2.13)}$$

where $\Sigma_1 = \mathbb{E}\big[\varphi_1\big(\mathcal{H}_1^{(i)};\beta_1^*\big)^{\otimes 2}\big]$.

By displays (C.2.12) and (C.2.13) we have that $-\sqrt{n}\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) + \sqrt{n}O(1)o_P\big(\big\|\hat{\beta}_1^{(n)} - \beta_1^*\big\|_2\big) = O_P(1)$. This implies that $\sqrt{n}\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) = O_P(1)$; thus, $\sqrt{n}O(1)o_P\big(\big\|\hat{\beta}_1^{(n)} - \beta_1^*\big\|_2\big) = o_P(1)$. Using these results, we have that

$$\sqrt{n}\big(\hat{\beta}_1^{(n)} - \beta_1^*\big) = -\sqrt{n}\big[\dot{\Phi}_1^*\big]^{-1}\Big\{\hat{\Phi}_1^{(n)}(\beta_1^*) - \Phi_1(\beta_1^*)\Big\} + o_P(1) = O_P(1).$$

Induction Step. For the induction step, for a given $t > 1$, we make the induction assumption that $\sqrt{n}\big(\hat{\beta}_{1:t-1}^{(n)} - \beta_{1:t-1}^*\big) = O_P(1)$. We then show that $\sqrt{n}\big(\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big) = O_P(1)$ and $\sqrt{n}\big\{\hat{\Phi}_{1:t}^{(n)}(\hat{\beta}_{1:t}^{(n)}) - \Phi_{1:t}(\hat{\beta}_{1:t}^{(n)})\big\} = \sqrt{n}\big\{\hat{\Phi}_{1:t}^{(n)}(\beta_{1:t}^*) - \Phi_{1:t}(\beta_{1:t}^*)\big\} + o_P(1)$.

Stochastic Equicontinuity Result. First, we will apply Lemma C.6.1 (Stochastic Equiconti-

nuity) to get that for any fixed vector $c \in \mathbb{R}^{d_{1:t}}$ (we use $d_{1:t} \triangleq \sum_{t'=1}^{t} d_{t'}$),

$$\sqrt{n}c^\top \left\{ \hat{\Phi}_{1:t}^{(n)}(\hat{\beta}_{1:t}^{(n)}) - \Phi_{1:t}(\hat{\beta}_{1:t}^{(n)}) \right\} = \sqrt{n}c^\top \left\{ \hat{\Phi}_{1:t}^{(n)}(\beta_{1:t}^*) - \Phi_{1:t}(\beta_{1:t}^*) \right\} + o_P(1). \quad \text{(C.2.14)}$$

We are able to apply Lemma C.6.1 (Stochastic Equicontinuity) because the following assumptions hold:

- Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) hold.
- $\hat{\beta}_{1:t-1}^{(n)} - \beta_{1:t-1}^* = O_P(1/\sqrt{n})$ by our induction assumption.
- Since assumption (NP1) (Finite Bracketing Integral) and Condition 5.3.3 (Lipschitz Policy Function) hold, we can apply Lemma C.1.5 (Finite Bracketing Integral) to get that for any vector $c \in \mathbb{R}^{d_t}$,

$$\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon,\ \left\{\pi_{2:t}(\,\cdot\,;\beta_{1:t-1})c^\top \varphi_t(\,\cdot\,;\beta_t)\right\}_{\beta_{1:t}\in B_{1:t}},\ L_{2+\alpha}(\mathcal{P}_{\pi^*})\right)} d\varepsilon < \infty,$$

  where $\pi_{2:t}(\,\cdot\,;\beta_{1:t-1}) \triangleq \prod_{t'=2}^{t} \pi_{t'}(\,\cdot\,;\beta_{t'-1})$; specifically see the result in Remark C.1.1 part (b2) below the statement of Lemma C.1.5.
- $\hat{\beta}_{1:t}^{(n)} \xrightarrow{P} \beta_{1:t}^*$ by Condition 5.3.1. Thus, by assumption (NP2) and continuous mapping theorem we have that for any vector $c \in \mathbb{R}^{d_t}$,

$$\nu\left(\pi_{2:t}(\,\cdot\,;\hat{\beta}^{(n)}_{1:t-1})c^\top\varphi_t(\,\cdot\,;\hat{\beta}^{(n)}_t),\ \pi_{2:t}(\,\cdot\,;\beta^*_{1:t-1})c^\top\varphi_t(\,\cdot\,;\beta^*_t)\right) \xrightarrow{P} 0, \text{ where}$$

$$\nu\left(\pi_{2:t}(\,\cdot\,;\beta_{1:t-1})c^\top\varphi_t(\,\cdot\,;\beta_t),\ \pi_{2:t}(\,\cdot\,;\beta'_{1:t-1})c^\top\varphi_t(\,\cdot\,;\beta'_t)\right)$$
$$\triangleq \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi_{2:t}(\,\cdot\,;\beta_{1:t-1})c^\top\varphi_t(\,\cdot\,;\beta_t) - \pi_{2:t}(\,\cdot\,;\beta'_{1:t-1})c^\top\varphi_t(\,\cdot\,;\beta'_t)\right\}^2\right]^{1/2}.$$

By display (C.2.14) and Cramer Wold device, we have that

$$\sqrt{n}\left\{\hat{\Phi}^{(n)}_{1:t}(\hat{\beta}^{(n)}_{1:t}) - \Phi_{1:t}(\hat{\beta}^{(n)}_{1:t})\right\} = \sqrt{n}\left\{\hat{\Phi}^{(n)}_{1:t}(\beta^*_{1:t}) - \Phi_{1:t}(\beta^*_{1:t})\right\} + o_P(1). \tag{C.2.15}$$

With the above result, for the induction step of the induction argument, we now just need to show that $\sqrt{n}(\hat{\beta}^{(n)}_{1:t} - \beta^*_{1:t}) = O_P(1)$.

Showing that $\sqrt{n}(\hat{\beta}^{(n)}_{1:t} - \beta^*_{1:t}) = O_P(1)$. By the definitions of $\hat{\beta}^{(n)}_{1:t}$ and $\beta^*_{1:t}$ from displays (5.3.9) and (5.3.7) respectively, we can rewrite the left-hand side of the display (C.2.15) above:

$$\sqrt{n}\big[\underbrace{\hat{\Phi}^{(n)}_{1:t}(\hat{\beta}^{(n)}_{1:t})}_{=o_P(1/\sqrt{n})} - \Phi_{1:t}(\hat{\beta}^{(n)}_{1:t})\big] = \sqrt{n}\big[\underbrace{\Phi_{1:t}(\beta^*_{1:t})}_{=0} - \Phi_{1:t}(\hat{\beta}^{(n)}_{1:t})\big] + o_P(1).$$

By Condition 5.5.1 (Differentiability of Policy Parameter Estimating Functions), the mapping $\beta_{1:t} \mapsto \Phi_{1:t}(\beta_{1:t})$ is differentiable at $\beta_{1:t} = \beta^*_{1:t}$ with the derivative matrix $\dot{\Phi}^*_{1:t} \triangleq \frac{\partial}{\partial \beta_{1:t}}\Phi_{1:t}(\beta_{1:t})\big|_{\beta_{1:t}=\beta^*_{1:t}}$. So,

$$= -\sqrt{n}\dot{\Phi}^*_{1:t}(\hat{\beta}^{(n)}_{1:t} - \beta^*_{1:t}) + \sqrt{n}o_P\big(\big\|\hat{\beta}^{(n)}_{1:t} - \beta^*_{1:t}\big\|_2\big) + o_P(1).$$

In summary, we have that

$$\sqrt{n}\left\{\hat{\Phi}_{1:t}^{(n)}(\beta_{1:t}^*) - \Phi_{1:t}(\beta_{1:t}^*)\right\} + o_P(1) = -\sqrt{n}\dot{\Phi}_{1:t}^*(\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*) + \sqrt{n}o_P\big(\big\|\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big\|_2\big).$$

By Condition 5.5.1, we can apply Lemma C.2.2 (Invertibility of $\dot{\Phi}_{1:t}^*$) to get that $\dot{\Phi}_{1:t}^*$ is invertible; so $\big[\dot{\Phi}_{1:t}^*\big]^{-1} = O(1)$ and

$$\begin{aligned}\sqrt{n}\big[\dot{\Phi}_{1:t}^*\big]^{-1}&\left\{\hat{\Phi}_{1:t}^{(n)}(\beta_{1:t}^*) - \Phi_{1:t}(\beta_{1:t}^*)\right\} + o_P(1)\\ &= -\sqrt{n}\big(\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big) + \sqrt{n}O(1)o_P\big(\big\|\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big\|_2\big). \quad \text{(C.2.16)}\end{aligned}$$

For now, we take as given that the following result in display (C.2.17) holds; we prove this at the end of this proof.

$$\sqrt{n}\big[\dot{\Phi}_{1:t}^*\big]^{-1}\left\{\hat{\Phi}_{1:t}^{(n)}(\beta_{1:t}^*) - \Phi_{1:t}(\beta_{1:t}^*)\right\} = O_P(1). \quad \text{(C.2.17)}$$

By displays (C.2.16) and (C.2.17) we have that $-\sqrt{n}\big(\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big) + \sqrt{n}O(1)o_P\big(\big\|\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big\|_2\big) = O_P(1)$. This implies that $\sqrt{n}\big(\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big) = O_P(1)$; thus, $\sqrt{n}O(1)o_P\big(\big\|\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big\|_2\big) = o_P(1)$. Using these results, we have that

$$\sqrt{n}\big(\hat{\beta}_{1:t}^{(n)} - \beta_{1:t}^*\big) = -\sqrt{n}\big[\dot{\Phi}_{1:t}^*\big]^{-1}\left\{\hat{\Phi}_{1:t}^{(n)}(\beta_{1:t}^*) - \Phi_{1:t}(\beta_{1:t}^*)\right\} + o_P(1) = O_P(1).$$

We now show that display (C.2.17) ***holds..*** For any fixed vector $c = [c_1, c_2, \cdots, c_t] \in$

$\mathbb{R}^{d_{1:t}}$,

$$\sqrt{n}c^\top \left\{ \hat{\Phi}^{(n)}_{1:t-1}(\beta^*_{1:t}) - \underbrace{\Phi_{1:t}(\beta^*_{1:t})}_{=0} \right\}$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{k=1}^{t} c_k^\top W^{(i)}_{2:k}(\beta^*_{1:k-1}, \hat{\beta}^{(n)}_{1:k-1}) \varphi_k(\mathcal{H}^{(i)}_k; \beta^*_k) \xrightarrow{D} \mathcal{N}\left(0, c^\top \Sigma_{1:t} c\right), \quad \text{(C.2.18)}$$

where

$$\Sigma_{1:t} \triangleq \mathbb{E}_{\pi^*_2} \left[\begin{pmatrix} \varphi_1(\mathcal{H}^{(i)}_1; \beta^*_1) \\ \varphi_2(\mathcal{H}^{(i)}_2; \beta^*_2) \\ \vdots \\ \varphi_t(\mathcal{H}^{(i)}_t; \beta^*_t) \end{pmatrix}^{\otimes 2} \right].$$

The asymptotic normality result above holds by Theorem C.5.1 (Weighted Martingale Triangular Array Central Limit Theorem). Specifically we can apply Theorem C.5.1 because:

- Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) hold.
- By our induction assumption, $\hat{\beta}^{(n)}_{t'} - \beta^*_{t'} = O_P(1/\sqrt{n})$ for all $t' \in [1\colon t-1]$.
- $\mathbb{E}_{\pi^*_{2:t}}\left[\left| \sum_{t'=1}^{t} \varphi_{t'}(\mathcal{H}^{(i)}_{t'}; \beta^*_{t'}) \right|^{2+\alpha} \right] < \infty$ by the Finite Bracketing Integral assumption (NP1).

Thus, by Cramer-Wold device and display (C.2.18) we have that

$$\sqrt{n}\left\{ \hat{\Phi}^{(n)}_{1:t}(\beta^*_{1:t}) - \Phi_{1:t}(\beta^*_{1:t}) \right\} \xrightarrow{D} \mathcal{N}(0, \Sigma_{1:t}).$$

By Condition 5.5.1, we can apply Lemma C.2.2 (Invertibility of $\dot{\Phi}^*_{1:t}$) to get that $\dot{\Phi}^*_{1:t}$ is

invertible. Thus, by continuous mapping theorem,

$$\sqrt{n}\left[\dot{\Phi}^{*}_{1:t}\right]^{-1}\left\{\hat{\Phi}^{(n)}_{1:t}(\beta^{*}_{1:t})-\Phi_{1:t}(\beta^{*}_{1:t})\right\}\xrightarrow{D}\mathcal{N}\left(0,\left[\dot{\Phi}^{*}_{1:t}\right]^{-1}\Sigma_{1:t}\left[\dot{\Phi}^{*}_{1:t}\right]^{-1,\top}\right).$$

The above implies that display (C.2.17) holds, i.e., that $\sqrt{n}\left[\dot{\Phi}^{*}_{1:t}\right]^{-1}\left\{\hat{\Phi}^{(n)}_{1:t}(\beta^{*}_{1:t})-\Phi_{1:t}(\beta^{*}_{1:t})\right\}$ $= O_P(1)$. ■

C.2.4 Invertibility of $\dot{\Phi}^{*}_{1:t}$ (Lemma C.2.2)

Lemma C.2.2 (Invertibility of $\dot{\Phi}^{*}_{1:t}$)**.** *Under Condition 5.5.1 (Differentiability of Policy Parameter Estimating Functions), for each $t \in [1\colon T-1]$, $\dot{\Phi}^{*}_{1:t}$ is invertible.*

In the proof of Lemma C.2.2 we will use the following proposition:

Proposition C.2.1 (Blockwise Inversion of Matrix)**.** *Let $A \in \mathbb{R}^{k\times k}$, $B \in \mathbb{R}^{k\times j}$, $C \in \mathbb{R}^{j\times k}$, and $D \in \mathbb{R}^{j\times j}$. If A and $D - CA^{-1}B$ are invertible then*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1} \end{bmatrix}.$$

Furthermore, in the special case that $B = 0$,

$$\begin{bmatrix} A & 0 \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ -D^{-1}CA^{-1} & D^{-1} \end{bmatrix}.$$

This is a result is proved in Proposition 3.9.7 of[10].

Proof of Lemma C.2.2. By Condition 5.5.1 (Differentiability of Policy Parameter

Estimating Functions), the mapping $\beta_{1:t} \mapsto \Phi_{1:t}(\beta_{1:t})$ is differentiable at $\beta_{1:t} = \beta^*_{1:t}$. Let $\dot{\Phi}^*_{1:t} \triangleq \frac{\partial}{\partial \beta_{1:t}} \Phi_t(\beta_{1:t})\big|_{\beta_{1:t}=\beta^*_{1:t}}$ Specifically,

$$\dot{\Phi}^*_{1:t} = \frac{\partial}{\partial \beta_{1:t}} \mathbb{E}_{\pi(\beta_{1:t-1})} \begin{bmatrix} \varphi_1(\mathcal{H}_1^{(i)}; \beta_1) \\ \varphi_2(\mathcal{H}_2^{(i)}; \beta_2) \\ \vdots \\ \varphi_t(\mathcal{H}_t^{(i)}; \beta_t) \end{bmatrix} \Bigg|_{\beta_{1:t}=\beta^*_{1:t}} = \begin{bmatrix} \dot{\Phi}^*_1 & 0 & 0 & \dots & 0 \\ V_{2,1} & \dot{\Phi}^*_2 & 0 & \dots & 0 \\ V_{3,1} & V_{3,2} & \dot{\Phi}^*_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ V_{t,1} & V_{t,2} & V_{t,3} & \dots & \dot{\Phi}^*_t \end{bmatrix},$$

where $V_{t,s} \triangleq \frac{\partial}{\partial \beta_s} \mathbb{E}_{\pi^*_{2:t}} \left[\varphi_t(\mathcal{H}_t^{(i)}; \beta^*_t) W_{s+1}^{(i)}(\beta_s, \beta^*_s) \right] \big|_{\beta_s=\beta^*_s} \in \mathbb{R}^{d_t \times d_s}$ and
$\dot{\Phi}^*_t \triangleq \frac{\partial}{\partial \beta_t} \Phi_t(\beta^*_{1:t-1}, \beta_t)\big|_{\beta_t=\beta^*_t} \triangleq \frac{\partial}{\partial \beta_t} \mathbb{E}_{\pi^*_{2:t}} \left[\varphi_t(\mathcal{H}_t^{(i)}; \beta_t) \right] \big|_{\beta_t=\beta^*_t}$.

Note that by repeatedly applying Proposition C.2.1 we can show that $\dot{\Phi}^*_{1:t}$ is invertible. Moreover, we will show that the inverse of $\dot{\Phi}^*_{1:t}$ is lower block triangular. To see this, consider the following induction argument:

- *Base case:* By Condition 5.5.1 (Differentiability of Policy Parameter Estimating Functions), $\dot{\Phi}^*_1$ and $\dot{\Phi}^*_2$ are both invertible. Thus, by Proposition C.2.1, the matrix $\begin{bmatrix} \dot{\Phi}^*_1 & 0 \\ V_{2,1} & \dot{\Phi}^*_2 \end{bmatrix}$ is invertible; moreover, its inverse is lower block triangular.
- *Induction step:* Suppose we know that for some $t \geq 2$, the inverse of the following

lower block triangular matrix is invertible:

$$\begin{bmatrix} \dot{\Phi}_1^* & 0 & 0 & \dots & 0 \\ V_{2,1} & \dot{\Phi}_2^* & 0 & \dots & 0 \\ V_{3,1} & V_{3,2} & \dot{\Phi}_3^* & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ V_{t-1,1} & V_{t-1,2} & V_{t-1,3} & \dots & \dot{\Phi}_{t-1}^* \end{bmatrix}. \tag{C.2.19}$$

By Condition 5.5.1 (Differentiability of Policy Parameter Estimating Functions), $\dot{\Phi}_t^*$ is invertible. By Proposition C.2.1 we can conclude that the following matrix is invertible and lower block triangular:

$$\begin{bmatrix} \dot{\Phi}_1^* & 0 & 0 & \dots & 0 & 0 \\ V_{2,1} & \dot{\Phi}_2^* & 0 & \dots & 0 & 0 \\ V_{3,1} & V_{3,2} & \dot{\Phi}_3^* & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ V_{t-1,1} & V_{t-1,2} & V_{t-1,3} & \dots & \dot{\Phi}_{t-1}^* & 0 \\ V_{t,1} & V_{t,2} & V_{t,3} & \dots & V_{t,t-1} & \dot{\Phi}_t^* \end{bmatrix}.$$

When applying Proposition C.2.1, take the matrix from display (C.2.19) (upper left-hand side of the matrix above) to be matrix A; take $\dot{\Phi}_t^*$ to be matrix D; take $\begin{bmatrix} V_{t,1} & V_{t,2} & V_{t,3} & \dots & V_{t,t-1} \end{bmatrix}$ to be matrix C; and take the block of zeros above $\dot{\Phi}_t^*$ to be matrix B.

By the above argument, $\dot{\Phi}_{1:t}^*$ is invertible for any $t \in [1 : T-1]$. ■

C.3 Main Asymptotic Results

Overview of Supplement C.3 Results.

- **Section C.3.1:** Consistency of $\hat{\theta}$ (Theorem 5.5.1)
- **Section C.3.2:** Equivalent Formulations for the Adaptive Sandwich Variance (Lemma C.3.1)
- **Section C.3.3:** Asymptotic Normality of $\hat{\theta}$ (Theorem 5.5.2)

C.3.1 Consistency of $\hat{\theta}$ (Theorem 5.5.1)

Proof of Theorem 5.5.1. Note that this argument is extremely similar to the proof for Theorem C.2.1. The proof will use the estimating functions $\Psi\left(\beta_{1:T-1}, \theta\right)$ and $\hat{\Psi}^{(n)}\left(\beta_{1:T-1}, \theta\right)$ defined earlier in displays (5.5.6) and (5.5.7) respectively.

For now, we take as given that display (C.2.4) below holds; we will show this result holds at the end of this proof.

$$\left\| \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \theta^*\right) - \Psi\left(\beta^*_{1:T-1}, \theta^*\right) \right\|_1 = o_p(1). \tag{C.3.1}$$

Let $\varepsilon > 0$. By assumption (C1), there exists some $\eta > 0$ such that if $\theta \in \mathbb{R}^{d_\theta}$ satisfies $\left\|\theta - \theta^*\right\|_1 > \varepsilon$, then $\left\|\Psi\left(\beta^*_{1:T-1}, \theta\right)\right\|_1 = \left\|\mathbb{E}_{\pi^*_{2:T}}\left[\psi(\mathcal{H}^{(i)}_T; \theta)\right]\right\|_1 > \eta > 0$. Thus,

$$\mathbb{P}\left(\left\|\hat{\theta} - \theta^*\right\|_1 > \varepsilon\right) \leq \mathbb{P}\left(\left\|\Psi\left(\beta^*_{1:T-1}, \hat{\theta}\right)\right\|_1 > \eta\right).$$

Note $\left\|\Psi\left(\beta^*_{1:T-1}, \hat{\theta}\right)\right\|_1 = \left\|\Psi\left(\beta^*_{1:T-1}, \hat{\theta}\right) - \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right) + \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1$
$\leq \left\|\Psi\left(\beta^*_{1:T-1}, \hat{\theta}\right) - \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 + \left\|\Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 = \left\|\Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 + o_P(1)$; the last equality holds by display (C.3.1). Thus,

$$\leq \mathbb{P}\left(\left\|\Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 > \eta - o_P(1)\right).$$

Note that $\Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n W^{(i)}_{2:T}\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\beta}^{(n)}_{1:T-1}\right)\psi\left(\mathcal{H}^{(i)}_T; \hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \psi\left(\mathcal{H}^{(i)}_T; \hat{\theta}\right)$
$= o_P(1/\sqrt{n}) = o_P(1)$; the second to last equality holds by the definition of $\hat{\theta}$ from display (5.3.2). Thus,

$$= \mathbb{P}\left(\left\|\hat{\Psi}^{(n)}\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right) - \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 > \eta - o_P(1)\right).$$

By assumption (C2), for any $\delta > 0$, there exists some $k < \infty$ such that
$\limsup_{n\to\infty} \mathbb{P}\left(\left\|\hat{\theta}\right\|_1 > k\right) \leq \delta$. We use this k below:

$$= \mathbb{P}\left(\left\|\hat{\Psi}^{(n)}\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right) - \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 \left\{\mathbb{I}_{\|\hat{\theta}\|_1 > k} + \mathbb{I}_{\|\hat{\theta}\|_1 \leq k}\right\} > \eta - o_P(1)\right)$$

$$\leq \mathbb{P}\left(\left\|\hat{\Psi}^{(n)}\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right) - \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 \mathbb{I}_{\|\hat{\theta}\|_1 \leq k} > \eta/2 - o_P(1)\right)$$
$$+ \mathbb{P}\left(\left\|\hat{\Psi}^{(n)}\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right) - \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 \mathbb{I}_{\|\hat{\theta}\|_1 > k} > \eta/2 - o_P(1)\right)$$

$$\leq \mathbb{P}\left(\left\|\hat{\Psi}^{(n)}\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right) - \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right)\right\|_1 \mathbb{I}_{\|\hat{\theta}\|_1 \leq k} > \eta/2 - o_P(1)\right)$$
$$+ \mathbb{P}\left(\left\|\hat{\theta}\right\|_1 > k\right) + o(1)$$

$$\leq \mathbb{P}\Bigg(\sup_{\theta \in \mathbb{R}^{d_\theta} \text{ s.t. } \|\theta\|_1 \leq k} \left\| \hat{\Psi}^{(n)}\left(\hat{\beta}^{(n)}_{1:T-1}, \theta\right) - \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \theta\right) \right\|_1 > \eta/2 - o_P(1) \Bigg)$$
$$+ \mathbb{P}\left(\left\| \hat{\theta} \right\|_1 > k \right) + o(1)$$

Since $\hat{\beta}^{(n)}_{1:T-1} \xrightarrow{P} \beta^*_{1:T-1}$ by Condition 5.3.1, thus $\mathbb{I}_{\hat{\beta}^{(n)}_{1:T-1} \in B_{1:T-1}} \xrightarrow{P} 1$; recall that $B_{1:T-1} \subset \mathbb{R}^{d_{1:T-1}}$ is a compact subset whose interior contains $\beta^*_{1:T-1}$. Thus,

$$\leq \underbrace{\mathbb{P}\Bigg(\sup_{\theta \in \mathbb{R}^{d_\theta} \text{ s.t. } \|\theta\|_1 \leq k} \sup_{\beta_{1:t-1} \in B_{1:t-1}} \left\| \hat{\Psi}^{(n)}\left(\beta_{1:T-1}, \theta\right) - \Psi\left(\beta_{1:T-1}, \theta\right) \right\|_1 > \eta/2 - o_P(1) \Bigg)}_{=o(1)}$$
$$+ \underbrace{\mathbb{P}\left(\left\| \hat{\theta} \right\|_1 > k \right)}_{\leq \delta} + o(1).$$

Note that the above converges to zero as $n \to \infty$ for the following reasons:

- By assumption (CP2), $\limsup_{n\to\infty} \mathbb{P}\left(\left\| \hat{\theta} \right\|_1 > k \right) \leq \delta$ and δ can be made arbitrarily small.

- Note that by Cramer Wold device, to show that

$$\sup_{\theta \in \mathbb{R}^{d_\theta} \text{ s.t. } \|\theta\|_1 \leq k} \sup_{\beta_{1:t-1} \in B_{1:t-1}} \left\| \hat{\Psi}^{(n)}\left(\beta_{1:T-1}, \theta\right) - \Psi\left(\beta_{1:T-1}, \theta\right) \right\|_1 \xrightarrow{P} 0,$$

  it is sufficient to show that for any vector $c \in \mathbb{R}^{d_\theta}$,

$$\sup_{\theta \in \mathbb{R}^{d_\theta} \text{ s.t. } \|\theta\|_1 \leq k} \sup_{\beta_{1:t-1} \in B_{1:t-1}} c^\top \left\{ \hat{\Psi}^{(n)}\left(\beta_{1:T-1}, \theta\right) - \Psi\left(\beta_{1:T-1}, \theta\right) \right\} \xrightarrow{P} 0.$$

Also note that

$$\sup_{\theta\in\mathbb{R}^{d_\theta} \text{ s.t. } \|\theta\|_1\le k} \sup_{\beta_{1:t-1}\in B_{1:t-1}} c^\top \left\{ \hat{\Psi}^{(n)}\big(\beta_{1:T-1},\theta\big) - \Psi\big(\beta_{1:T-1},\theta\big)\right\}$$

$$= \sup_{\theta\in\mathbb{R}^{d_\theta} \text{ s.t. } \|\theta\|_1\le k} \sup_{\beta_{1:t-1}\in B_{1:t-1}} \frac{1}{n}\sum_{i=1}^{n} \Bigg\{ W^{(i)}_{2:T}(\beta_{1:T-1},\hat{\beta}^{(n)}_{1:T-1}) c^\top \psi(\mathcal{H}^{(i)}_T;\theta) - \mathbb{E}\left[W^{(i)}_{2:T}(\beta_{1:T-1},\hat{\beta}^{(n)}_{1:T-1}) c^\top \psi(\mathcal{H}^{(i)}_T;\theta)\right]\Bigg\} \xrightarrow{P} 0.$$

The above convergence result holds by Theorem C.4.2 (Weighted Martingale Triangular Array Uniform Weak Law of Large Numbers). Specifically we are able to apply Theorem C.4.2 because Condition 5.3.2 holds and
$N_{[\,]}\left(\varepsilon, \mathcal{F}_{\Pi c^\top \psi}(B_{1:T-1},k), L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ for any $c\in\mathbb{R}^{d_\theta}$ and any $\varepsilon>0$, where

$$\mathcal{F}_{\Pi c^\top\psi}(B_{1:T-1},k) \triangleq \left\{ \left[\prod_{t=2}^{T} \pi_t(\,\cdot\,;\beta_{t-1})\right] c^\top \psi(\,\cdot\,;\theta)\right\}_{\beta_{1:T-1}\in B_{1:T-1},\ \theta\in\mathbb{R}^{d_\theta} \text{ s.t. } \|\theta\|_1\le k}.$$

The above finite bracketing number result holds since using assumption (CP3) and Condition 5.3.3, we can apply by Lemma C.1.5 (specifically see Remark C.1.1 part (a1)).

We now show that display (C.3.1) ***holds.*** Let $\beta_{1:T-1}\in B_{1:T-1}$.

$$\left\|\Psi\big(\beta_{1:T-1},\theta^*\big) - \Psi\big(\beta^*_{1:T-1},\theta^*\big)\right\|_1$$

$$= \left\| \mathbb{E}\left[\left\{ \prod_{t=2}^{T} W_t^{(i)}\left(\beta_{t-1}, \hat{\beta}_{t-1}^{(n)}\right) - W_t^{(i)}\left(\beta_{t-1}^*, \hat{\beta}_{t-1}^{(n)}\right) \right\} \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right] \right\|_1$$

$$= \left\| \mathbb{E}_{\pi_{2:T}^*}\left[\left\{ \prod_{t=2}^{T} W_t^{(i)}\left(\beta_{t-1}, \beta_{t-1}^*\right) - W_t^{(i)}\left(\beta_{t-1}^*, \beta_{t-1}^*\right) \right\} \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right] \right\|_1$$

By Jensen's inequality,

$$\leq \mathbb{E}_{\pi_{2:T}^*}\left[\left| \prod_{t=2}^{T} W_t^{(i)}\left(\beta_{t-1}, \beta_{t-1}^*\right) - W_t^{(i)}\left(\beta_{t-1}^*, \beta_{t-1}^*\right) \right| \left\| \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right\|_1 \right]$$

By definition of $W_t^{(i)}\left(\beta_{t-1}, \beta_{t-1}^*\right)$ from display (5.5.5),

$$= \mathbb{E}_{\pi_{2:T}^*}\left[\left| \prod_{t=2}^{T} \pi_t\left(A_t^{(i)}, S_t^{(i)}; \beta_{t-1}\right) - \prod_{t=2}^{T} \pi_t\left(A_t^{(i)}, S_t^{(i)}; \beta_{t-1}^*\right) \right| \left\{ \prod_{t=2}^{T} \frac{1}{\pi_t^*\left(A_t^{(i)}, S_t^{(i)}\right)} \right\} \left\| \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right\|_1 \right]$$

By Condition 5.3.2 (Minimum Exploration), $\pi_t^*\left(A_t^{(i)}, S_t^{(i)}\right)^{-1} \leq \pi_{\min}^{-1}$ a.s. Thus,

$$\leq \pi_{\min}^{-(T-1)} \mathbb{E}_{\pi_{2:T}^*}\left[\left| \prod_{t=2}^{T} \pi_t\left(A_t^{(i)}, S_t^{(i)}; \beta_{t-1}\right) - \prod_{t=2}^{T} \pi_t\left(A_t^{(i)}, S_t^{(i)}; \beta_{t-1}^*\right) \right| \left\| \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right\|_1 \right]$$

By Condition 5.3.3 and Lemma C.1.4 (Product of Lipschitz Policy Functions are Lipschitz),

$$\leq \pi_{\min}^{-(T-1)} \mathbb{E}_{\pi_{2:T}^*}\left[\left\| \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right\|_1 \left\{ \sum_{t=2}^{T} \dot{\pi}_t\left(A_t^{(i)}, S_t^{(i)}\right) \right\} \left\| \beta_{1:T-1} - \beta_{1:T-1}^* \right\|_2 \right]$$

By linearity of expectations,

$$= \pi_{\min}^{-(T-1)} \left\{ \sum_{t=2}^{T} \mathbb{E}_{\pi^*_{2:T}} \left[\left\| \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right\|_1 \dot{\pi}_t(A_t^{(i)}, S_t^{(i)}) \right] \right\} \left\| \beta_{1:T-1} - \beta^*_{1:T-1} \right\|_2.$$

Thus, by consolidating the above results, we have that

$$\begin{aligned} &\left\| \Psi\left(\hat{\beta}^{(n)}_{1:T-1}, \theta^*\right) - \Psi\left(\beta^*_{1:T-1}, \theta^*\right) \right\|_1 \\ &\quad \leq \pi_{\min}^{-(T-1)} \left\{ \sum_{t=2}^{T} \mathbb{E}_{\pi^*_{2:T}} \left[\left\| \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right\|_1 \dot{\pi}_t(A_t^{(i)}, S_t^{(i)}) \right] \right\} \left\| \hat{\beta}^{(n)}_{1:T-1} - \beta^*_{1:T-1} \right\|_2 = o_P(1). \end{aligned}$$

The last limit above holds because

- $\left\| \hat{\beta}^{(n)}_{1:T-1} - \beta^*_{1:T-1} \right\|_2 = o_P(1)$ since $\hat{\beta}^{(n)}_{1:T-1} \xrightarrow{P} \beta^*_{1:T-1}$ by Condition 5.3.1.
- By assumption (C3) (Finite Bracketing Number for Policy Functions), there exists a function F_ψ such that $\left\| \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right\|_1 \leq F_\psi(\mathcal{H}_t^{(i)})$ a.s. and for all $t \in [2\colon T]$, $\mathbb{E}_{\pi^*_{2:T}}\left[F_\psi\left(\mathcal{H}_T^{(i)}\right) \dot{\pi}_t(A_t^{(i)}, S_t^{(i)})\right] < \infty$. Thus,

$$\mathbb{E}_{\pi^*_{2:T}} \left[\left\| \psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) \right\|_1 \dot{\pi}_t(A_t^{(i)}, S_t^{(i)}) \right] \leq \mathbb{E}_{\pi^*_{2:T}} \left[F_\psi\left(\mathcal{H}_T^{(i)}\right) \dot{\pi}_t(A_t^{(i)}, S_t^{(i)}) \right] < \infty.$$

We have now shown that display (C.3.1) holds. ■

C.3.2 Equivalent Formulations for the Adaptive Sandwich Variance (Lemma C.3.1)

Lemma C.3.1 (Equivalent Formulations for the Adaptive Sandwich Variance)**.** *Let Condition 5.5.1 (Differentiability of Policy Parameter Estimating Functions) and assumptions*

(N1) and (N2) from Theorem 5.5.2 (Asymptotic Normality hold). Also let $\Sigma_{1:T}$ *as defined in display* (5.5.11) *be finite.*

Then, the lower-right $d_\theta \times d_\theta$ *block of limiting variance from display* (5.5.10)*, i.e.,*

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} \Sigma_{1:T} \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1,\top}, \tag{C.3.2}$$

equals the adaptive sandwich variance, i.e., $[\dot{\Psi}^*]^{-1}\Sigma^{\text{adapt}}[\dot{\Psi}^*]^{-1,\top}$ *from display* (5.5.3)*.*

Proof of Lemma C.3.1. By the definition of Σ^{adapt} from display (5.5.3), it is sufficient to show that the lower-right $d_\theta \times d_\theta$ block of the the limiting variance from display (C.3.2) above equals the following

$$[\dot{\Psi}^*]^{-1}\Sigma^{\text{adapt}}[\dot{\Psi}^*]^{-1,\top} = [\dot{\Psi}^*]^{-1}\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\psi\big(\mathcal{H}_T^{(i)};\theta^*\big) + \dot{\Psi}^* \sum_{t=1}^{T-1} M_t \varphi_t\big(\mathcal{H}_t^{(i)};\beta_t^*\big)\right\}^{\otimes 2}\right][\dot{\Psi}^*]^{-1,\top}.$$

Consider the following matrix from display (C.3.2) (the terms in the matrix below are derivatives that exist by Condition 5.5.1, and assumptions (N1) and (N2)):

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}. \tag{C.3.3}$$

By Proposition C.2.1 (Blockwise Inversion of Matrix), we have that for a block matrix $\begin{bmatrix} A & 0 \\ C & D \end{bmatrix}$, if square matrices A and D are invertible, then the whole matrix is invertible

and

$$\begin{bmatrix} A & 0 \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ -D^{-1}CA^{-1} & D^{-1} \end{bmatrix}.$$

By Condition 5.5.1 and Lemma C.2.2 (Invertibility of $\dot{\Phi}^*_{1:t}$), $\dot{\Phi}^*_{1:t}$ is invertible and $\dot{\Psi}^*$ is invertible by assumption (N1). Thus the matrix from display (C.3.3) is invertible and

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} = \begin{bmatrix} \left\{\dot{\Phi}^*_{1:T-1}\right\}^{-1} & 0 \\ -\left\{\dot{\Psi}^*\right\}^{-1} V_{T,1:T-1} \left\{\dot{\Phi}^*_{1:T-1}\right\}^{-1} & \left\{\dot{\Psi}^*\right\}^{-1} \end{bmatrix}.$$

Recall in display (5.5.32) we defined the following matrices:

$$M_{1:T-1} \triangleq \left[M_1, M_2, \ldots, M_{T-1}\right] \triangleq -\left\{\dot{\Psi}^*\right\}^{-1} V_{T,1:T-1} \left\{\dot{\Phi}^*_{1:T-1}\right\}^{-1} \in \mathbb{R}^{d_\vartheta \times d_{1:T-1}}. \quad \text{(C.3.4)}$$

Above we use $d_{1:T-1} \triangleq \sum_{t=1}^{T-1} d_t$. Thus,

$$\begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} \Sigma_{1:T} \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1,\top}$$

$$= \begin{bmatrix} \left\{\dot{\Phi}^*_{1:T-1}\right\}^{-1} & 0 \\ M_{1:T-1} & \left\{\dot{\Psi}^*\right\}^{-1} \end{bmatrix} \begin{bmatrix} \Sigma_{1:T-1} & U_{1:T-1,T} \\ U_{T,1:T-1} & \Sigma \end{bmatrix} \begin{bmatrix} \left\{\dot{\Phi}^*_{1:T-1}\right\}^{-1} & 0 \\ M_{1:T-1} & \left\{\dot{\Psi}^*\right\}^{-1} \end{bmatrix}^{\top}. \quad \text{(C.3.5)}$$

Above $\Sigma \triangleq \mathbb{E}_{\pi^*_{2:T-1}}\left[\psi(\mathcal{H}^{(i)}_T; \theta^*)^{\otimes 2}\right]$, $\Sigma_{1:T-1} \triangleq \mathbb{E}_{\pi^*_{2:T-1}}\left[\varphi_{1:T-1}(\mathcal{H}^{(i)}_{T-1}; \beta^*_{1:T-1})^{\otimes 2}\right]$,

$U_{T,1:T-1} \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\psi(\mathcal{H}_T^{(i)};\theta^*)\varphi_{1:T-1}(\mathcal{H}_{T-1}^{(i)};\beta^*_{1:T-1})^\top\right]$, and $U_{1:T-1,T} \triangleq U_{T,1:T-1}^\top$, where

$$\varphi_{1:T-1}(\mathcal{H}_{T-1}^{(i)};\beta^*_{1:T-1}) \triangleq \begin{pmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1^*) \\ \varphi_2(\mathcal{H}_2^{(i)};\beta_2^*) \\ \vdots \\ \varphi_{T-1}(\mathcal{H}_{T-1}^{(i)};\beta^*_{T-1}) \end{pmatrix}. \tag{C.3.6}$$

Thus display (C.3.5) equals the following:

$$= \begin{bmatrix} \left\{\dot\Phi^*_{1:T-1}\right\}^{-1}\Sigma_{1:T-1} & \left\{\dot\Phi^*_{1:T-1}\right\}^{-1}U_{1:T-1,T} \\ M_{1:T-1}\Sigma_{1:T-1} + \left\{\dot\Psi^*\right\}^{-1}U_{T,1:T-1} & M_{1:T-1}U_{1:T-1,T} + \left\{\dot\Psi^*\right\}^{-1}\Sigma \end{bmatrix} \begin{bmatrix} \left\{\dot\Phi^*_{1:T-1}\right\}^{-1,\top} & M_{1:T-1}^\top \\ 0 & \left\{\dot\Psi^*\right\}^{-1,\top} \end{bmatrix}$$

Thus, the lower-right $d_\theta \times d_\theta$ block of the product of matrices above equals the following:

$$M_{1:T-1}\Sigma_{1:T-1}M_{1:T-1}^\top + \left\{\dot\Psi^*\right\}^{-1}U_{T,1:T-1}M_{1:T-1}^\top + M_{1:T-1}U_{1:T-1,T}\left\{\dot\Psi^*\right\}^{-1,\top} + \left\{\dot\Psi^*\right\}^{-1}\Sigma\left\{\dot\Psi^*\right\}^{-1,\top}$$

$$= \left\{\dot\Psi^*\right\}^{-1}\Big[\dot\Psi^* M_{1:T-1}\Sigma_{1:T-1}M_{1:T-1}^\top\left\{\dot\Psi^*\right\}^\top + U_{T,1:T-1}M_{1:T-1}^\top\left\{\dot\Psi^*\right\}^\top + \dot\Psi^* M_{1:T-1}U_{1:T-1,T} + \Sigma\Big]\left\{\dot\Psi^*\right\}^{-1,\top}$$

$$= \left\{\dot{\Psi}^*\right\}^{-1} \mathbb{E}_{\pi^*_{2:T}}\left[\left\{\psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) + \dot{\Psi}^* M_{1:T-1} \varphi_{1:T-1}\left(\mathcal{H}_{T-1}^{(i)}; \beta^*_{1:T-1}\right)\right\}^{\otimes 2}\right] \left\{\dot{\Psi}^*\right\}^{-1,\top}$$

By the definition of $M_{1:T-1}$ from display (C.3.4) and the definition of $\varphi_{1:T-1}$ from display (C.3.6),

$$= \left\{\dot{\Psi}^*\right\}^{-1} \mathbb{E}_{\pi^*_{2:T}}\left[\left\{\psi\left(\mathcal{H}_T^{(i)}; \theta^*\right) + \dot{\Psi}^* \sum_{t=1}^{T-1} M_t \varphi_t\left(\mathcal{H}_t^{(i)}; \beta_t^*\right)\right\}^{\otimes 2}\right] \left\{\dot{\Psi}^*\right\}^{-1,\top}.$$

We have now shown the desired result. ■

C.3.3 Asymptotic Normality of $\hat{\theta}$ (Theorem 5.5.2)

Proof of Theorem 5.5.2. By Lemma C.3.1 (Equivalent Formulations for the Adaptive Sandwich Variance) above, it is sufficient to show that

$$\sqrt{n}\begin{pmatrix} \hat{\beta}^{(n)}_{1:T-1} - \beta^*_{1:T-1} \\ \hat{\theta} - \theta^* \end{pmatrix} \xrightarrow{D} \mathcal{N}\left(0, \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} \Sigma_{1:T} \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1,\top}\right).$$

This proof will use the estimating functions $\Psi\left(\beta_{1:T-1}, \theta\right)$, $\hat{\Psi}^{(n)}\left(\beta_{1:T-1}, \theta\right)$, $\Phi_{1:T-1}(\beta_{1:T-1})$, and $\hat{\Phi}^{(n)}_{1:T-1}(\beta_{1:T-1})$ defined earlier in displays (5.5.6), (5.5.7), (5.5.8), and (5.5.9) respectively.

We now state several equalities and discuss why they hold below:

$$
\begin{aligned}
&-\sqrt{n}\begin{bmatrix}\hat{\Phi}^{(n)}_{1:T-1}\big(\hat{\beta}^{(n)}_{1:T-1}\big) - \Phi_{1:T-1}\big(\hat{\beta}^{(n)}_{1:T-1}\big)\\ \underbrace{\hat{\Psi}^{(n)}\big(\hat{\beta}^{(n)}_{1:T-1},\hat{\theta}\big)}_{=o_P(1/\sqrt{n})} - \Psi\big(\hat{\beta}^{(n)}_{1:T-1},\hat{\theta}\big)\end{bmatrix}\\
&\qquad\underbrace{=}_{(a)} -\sqrt{n}\begin{bmatrix}\Phi_{1:T-1}\big(\beta^{*}_{1:T-1}\big) - \Phi_{1:T-1}\big(\hat{\beta}^{(n)}_{1:T-1}\big)\\ \underbrace{\Psi\big(\beta^{*}_{1:T-1},\theta^{*}\big)}_{=0} - \Psi\big(\hat{\beta}^{(n)}_{1:T-1},\hat{\theta}\big)\end{bmatrix} + o_P(1)\\
&\underbrace{=}_{(b)}\sqrt{n}\begin{bmatrix}\dot{\Phi}^{*}_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^{*}\end{bmatrix}\begin{bmatrix}\hat{\beta}^{(n)}_{1:T-1}-\beta^{*}_{1:T-1}\\ \hat{\theta}-\theta^{*}\end{bmatrix} + \sqrt{n}o_P\left(\left\|\begin{matrix}\hat{\beta}^{(n)}_{1:T-1}-\beta^{*}_{1:T-1}\\ \hat{\theta}-\theta^{*}\end{matrix}\right\|_2\right) + o_P(1).
\end{aligned}
\tag{C.3.7}
$$

Equality (a). Equality (a) above holds since $\hat{\Psi}^{(n)}\big(\hat{\beta}^{(n)}_{1:T-1},\hat{\theta}\big) = o_P(1/\sqrt{n})$ and $\Psi\big(\beta^{*}_{1:T-1},\theta^{*}\big) = 0$ by the definitions of $\hat{\theta}$ and θ^{*} from displays (5.3.2) and (5.3.1) respectively; also since $\hat{\Phi}^{(n)}_{1:T-1}\big(\hat{\beta}^{(n)}_{1:T-1}\big) = o_P(1/\sqrt{n})$ and $\Phi_{1:T-1}\big(\beta^{*}_{1:T-1}\big) = 0$ by the definitions of $\hat{\beta}^{(n)}_{t}$ and β^{*}_{t} from displays (5.3.9) and (5.3.7).

Equality (b). Equality (b) above holds by a Taylor series expansion. Specifically by assumptions (N1) and (N2), the mapping $\big(\beta_{1:T-1},\theta\big) \mapsto \Psi(\beta_{1:T-1},\theta)$ is differentiable at $\big(\beta_{1:T-1},\theta\big) = \big(\beta^{*}_{1:T-1},\theta^{*}\big)$. Additionally the mapping $\beta_{1:T-1} \mapsto \Phi_{1:T-1}(\beta_{1:T-1})$ is differen-

tiable at $\beta_{1:T-1} = \beta^*_{1:T-1}$ by Condition 5.5.1. Thus,

$$\frac{\partial}{\partial(\beta_{1:T-1},\theta)}\begin{bmatrix}\Phi_{1:T-1}(\beta_{1:T-1})\\ \Psi(\beta_{1:T-1},\theta)\end{bmatrix}\Big|_{(\beta_{1:T-1},\theta)=(\beta^*_{1:T-1},\theta^*)}$$
$$=\begin{bmatrix}\frac{\partial}{\partial\beta_{1:T-1}}\Phi_{1:T-1}(\beta_{1:T-1})\big|_{\beta_{1:T-1}=\beta^*_{1:T-1}} & \frac{\partial}{\partial\theta}\Phi_{1:T-1}(\beta^*_{1:T-1})\big|_{\theta=\theta^*}\\ \frac{\partial}{\partial\beta_{1:T-1}}\Psi(\beta_{1:T-1},\theta^*)\big|_{\beta_{1:T-1}=\beta^*_{1:T-1}} & \frac{\partial}{\partial\theta}\Psi(\beta^*_{1:T-1},\theta)\big|_{\theta=\theta^*}\end{bmatrix}=\begin{bmatrix}\dot{\Phi}^*_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^*\end{bmatrix}.$$

As mentioned below display (5.5.12), $\frac{\partial}{\partial\theta}\Phi_{1:T-1}(\beta^*_{1:T-1})\big|_{\theta=\theta^*} = 0$ since $\Phi_{1:T-1}(\beta^*_{1:T-1})$ is not a function of θ.

We now state the next set of results and discuss why they hold below:

$$-\sqrt{n}\begin{bmatrix}\hat{\Phi}^{(n)}_{1:T-1}(\hat{\beta}^{(n)}_{1:T-1}) - \Phi_{1:T-1}(\hat{\beta}^{(n)}_{1:T-1})\\ \hat{\Psi}^{(n)}(\hat{\beta}^{(n)}_{1:T-1},\hat{\theta}) - \Psi(\hat{\beta}^{(n)}_{1:T-1},\hat{\theta})\end{bmatrix}$$
$$\underbrace{=}_{(c)} -\sqrt{n}\begin{bmatrix}\hat{\Phi}^{(n)}_{1:T-1}(\beta^*_{1:T-1}) - \Phi_{1:T-1}(\beta^*_{1:T-1})\\ \hat{\Psi}^{(n)}(\beta^*_{1:T-1},\theta^*) - \Psi(\beta^*_{1:T-1},\theta^*)\end{bmatrix} + o_P(1)\underbrace{\xrightarrow{D}}_{(d)}\mathcal{N}(0,\Sigma_{1:T}). \quad \text{(C.3.8)}$$

Equality (c). Equality (c) above is an asymptotic equicontinuity result. We now discuss why equality (c) holds. First note that by Conditions 5.3.1-5.5.2 and Remark C.2.1 (which shows that Condition 5.5.2 implies that assumptions (NP1) and (NP2) hold), we can apply Theorem C.2.2 (Asymptotic Equicontinuity for Policy Parameters) to get the following results:

$$\hat{\beta}^{(n)}_{1:T-1} - \beta^*_{1:T-1} = O_P(1/\sqrt{n}) \quad \text{(C.3.9)}$$

and

$$\sqrt{n}\left\{\hat{\Phi}^{(n)}_{1:T-1}(\hat{\beta}^{(n)}_{1:T-1}) - \Phi_{1:T-1}(\hat{\beta}^{(n)}_{1:T-1})\right\}$$
$$= \sqrt{n}\left\{\hat{\Phi}^{(n)}_{1:T-1}(\beta^*_{1:T-1}) - \Phi_{1:T-1}(\beta^*_{1:T-1})\right\} + o_P(1). \quad \text{(C.3.10)}$$

We now apply Lemma C.6.1 (Stochastic Equicontinuity) to get that

$$\sqrt{n}\left\{\hat{\Psi}^{(n)}(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}) - \Psi(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta})\right\}$$
$$= \sqrt{n}\left\{\hat{\Psi}^{(n)}(\beta^*_{1:T-1}, \theta^*) - \Psi(\beta^*_{1:T-1}, \theta^*)\right\} + o_P(1), \quad \text{(C.3.11)}$$

We are able to apply Lemma C.6.1 (Stochastic Equicontinuity) because the following assumptions hold:

- Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) hold.
- Note that $\hat{\beta}^{(n)}_{1:T-1} - \beta^*_{1:T-1} = O_P(1/\sqrt{n})$ by display (C.3.9).
- Since assumption (N4) (Finite Bracketing Integral) and Condition 5.3.3 (Lipschitz Policy Function) hold, we can apply Lemma C.1.5 (specifically see Remark C.1.1 part (a2)) to get that for any $c \in \mathbb{R}^{d_\vartheta}$,

  $$\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon,\ \left\{\pi_{2:T}(\,\cdot\,;\beta_{1:T-1})c^\top \psi(\,\cdot\,;\theta)\right\}_{\beta_{1:T-1}\in B_{1:T-1},\theta\in\Theta},\ L_{2+\alpha}(\mathcal{P}_{\pi^*})\right)}\, d\varepsilon < \infty,$$

  where $\pi_{2:T}(\,\cdot\,;\beta_{1:T-1}) \triangleq \prod_{t'=2}^{T} \pi_{t'}(\,\cdot\,;\beta_{t'-1})$.

- $\left(\hat{\beta}^{(n)}_{1:T-1}, \hat{\theta}\right) \xrightarrow{P} \left(\beta^*_{1:T-1}, \theta^*\right)$ by Slutsky's theorem since $\hat{\beta}^{(n)}_{1:T-1} \xrightarrow{P} \beta^*_{1:T-1}$ by Condition 5.3.1 and since $\hat{\theta} \xrightarrow{P} \theta^*$ by assumption. Thus, by assumption (N3) and continuous mapping theorem we have that for any $c \in \mathbb{R}^{d_\theta}$,
$\nu\left(c^\top \pi_{2:T}(\,\cdot\,; \hat{\beta}^{(n)}_{1:T-1})\psi(\,\cdot\,;\hat{\theta}),\ c^\top \pi_{2:T}(\,\cdot\,;\beta^*_{1:T-1})\psi(\,\cdot\,;\theta^*)\right) \xrightarrow{P} 0$, where

$$\nu\Big(\pi_{2:T}(\,\cdot\,;\beta_{1:T-1})c^\top\psi(\,\cdot\,;\theta),\ \pi_{2:T}(\,\cdot\,;\beta'_{1:T-1})c^\top\psi(\,\cdot\,;\theta')\Big) \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi_{2:T}(\,\cdot\,;\beta_{1:T-1})c^\top\psi(\,\cdot\,;\theta) - \pi_{2:T}(\,\cdot\,;\beta'_{1:T-1})c^\top\psi(\,\cdot\,;\theta')\right\}^2\right]^{1/2}.$$

Equality (c) holds since by Slutsky's Theorem, we can combine the results above from displays (C.3.10) and (C.3.11).

Asymptotic normality result (d). Equality (d) above holds by Theorem C.5.1 (Weighted Martingale Triangular Array Central Limit Theorem). Specifically note that for any fixed vector $c = [c_1, c_2, \ldots, c_T] \in \mathbb{R}^{d_{1:T-1}+d_\theta}$,

$$-\sqrt{n}c^\top \begin{bmatrix} \hat{\Phi}^{(n)}_{1:T-1}\left(\beta^*_{1:T-1}\right) - \Phi_{1:T-1}\left(\beta^*_{1:T-1}\right) \\ \hat{\Psi}^{(n)}\left(\beta^*_{1:T-1}, \theta^*\right) - \Psi\left(\beta^*_{1:T-1}, \theta^*\right) \end{bmatrix}$$

$$= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T-1} c_t^\top W^{(i)}_{2:t}(\beta^*_{1:t-1}, \hat{\beta}^{(n)}_{1:t-1})\varphi_t(\mathcal{H}^{(i)}_t; \beta^*_t) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} c_T^\top W^{(i)}_{2:T}(\beta^*_{1:T-1}, \hat{\beta}^{(n)}_{1:T-1})\psi(\mathcal{H}^{(i)}_T; \theta^*) \xrightarrow{D} \mathcal{N}\left(0, c^\top\Sigma_{1:T}c\right) \quad \text{(C.3.12)}$$

where

$$\Sigma_{1:T} \triangleq \mathbb{E}_{\pi^*_{2:T}}\left[\begin{pmatrix} \varphi_1(\mathcal{H}_1^{(i)};\beta_1^*) \\ \varphi_2(\mathcal{H}_2^{(i)};\beta_2^*) \\ \vdots \\ \varphi_{T-1}(\mathcal{H}_{T-1}^{(i)};\beta_{T-1}^*) \\ \psi(\mathcal{H}_T^{(i)};\theta^*) \end{pmatrix}^{\otimes 2}\right].$$

The final asymptotic normality result above holds by Theorem C.5.1 (Weighted Martingale Triangular Array Central Limit Theorem). Specifically we can apply Theorem C.5.1 because:

- Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) hold.
- $\hat{\beta}_{t'}^{(n)} - \beta_{t'}^* = O_P(1/\sqrt{n})$ for all $t' \in [1\colon T-1]$ by display (C.3.9).
- $\mathbb{E}_{\pi^*_{2:T}}\left[\left|c_T^\top \psi(\mathcal{H}_T^{(i)};\theta^*)\right|^{2+\alpha}\right] < \infty$ by Finite Bracketing Integral assumption (N4). Also, $\mathbb{E}_{\pi^*_{2:t}}\left[\left|c_t^\top \varphi_t(\mathcal{H}_t^{(i)};\beta_t^*)\right|^{2+\alpha}\right] < \infty$ for each $t \in [1\colon T-1]$ by Condition 5.5.2.

Thus, by Cramer-Wold device and display (C.3.12) we have that

$$-\sqrt{n}\begin{bmatrix} \hat{\Phi}_{1:T-1}^{(n)}(\beta_{1:T-1}^*) - \Phi_{1:T-1}(\beta_{1:T-1}^*) \\ \hat{\Psi}^{(n)}(\beta_{1:T-1}^*,\theta^*) - \Psi(\beta_{1:T-1}^*,\theta^*) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0,\Sigma_{1:T}). \tag{C.3.13}$$

Thus, we have shown that asymptotic normality result (d) holds.

Consolidating Results. By consolidating the results from displays (C.3.7) and (C.3.8)

above, and applying Slutsky's theorem we get the following result:

$$\sqrt{n}\begin{bmatrix}\dot{\Phi}^*_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^*\end{bmatrix}\begin{bmatrix}\hat{\beta}^{(n)}_{1:T-1}-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{bmatrix}+\sqrt{n}o_P\left(\left\|\begin{matrix}\hat{\beta}^{(n)}_{1:T-1}-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{matrix}\right\|_2\right)\xrightarrow{D}\mathcal{N}\left(0,\Sigma_{1:T}\right). \tag{C.3.14}$$

Note that $\dot{\Psi}^*$ is invertible by assumption (N1) and $\dot{\Phi}^*_{1:T-1}$ is invertible by Condition 5.5.1 and Lemma C.2.2. By Proposition C.2.1 (Block Inversion of Matrices), this is sufficient for $\begin{bmatrix}\dot{\Phi}^*_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^*\end{bmatrix}$ to be invertible. Thus, by continuous mapping theorem, and display (C.3.14),

$$\begin{aligned}\sqrt{n}\begin{bmatrix}\hat{\beta}^{(n)}_{1:T-1}-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{bmatrix}+\sqrt{n}O(1)o_P\left(\left\|\begin{matrix}\hat{\beta}^{(n)}_{1:T-1}-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{matrix}\right\|_2\right)+o_P(1)\\ \xrightarrow{D}\mathcal{N}\left(0,\begin{bmatrix}\dot{\Phi}^*_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^*\end{bmatrix}^{-1}\Sigma_{1:T}\begin{bmatrix}\dot{\Phi}^*_{1:T-1} & 0\\ V_{T,1:T-1} & \dot{\Psi}^*\end{bmatrix}^{-1,\top}\right).\end{aligned} \tag{C.3.15}$$

The asymptotic normality result above in display (C.3.15) implies that

$$\sqrt{n}\begin{bmatrix}\hat{\beta}^{(n)}_{1:T-1}-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{bmatrix}+\sqrt{n}O(1)o_P\left(\left\|\begin{matrix}\hat{\beta}^{(n)}_{1:T-1}-\beta^*_{1:T-1}\\ \hat{\theta}-\theta^*\end{matrix}\right\|_2\right)=O_P(1).$$

This implies that $\sqrt{n}\begin{bmatrix} \hat{\beta}^{(n)}_{1:T-1} - \beta^*_{1:T-1} \\ \hat{\theta} - \theta^* \end{bmatrix} = O_P(1)$. Thus, we have that

$$\sqrt{n}O(1)o_P\left(\left\|\begin{bmatrix} \hat{\beta}^{(n)}_{1:T-1} - \beta^*_{1:T-1} \\ \hat{\theta} - \theta^* \end{bmatrix}\right\|_2\right) = o_P(1).$$

So, by display (C.3.15) and Slutsky's Theorem we have that

$$\sqrt{n}\begin{bmatrix} \hat{\beta}^{(n)}_{1:T-1} - \beta^*_{1:T-1} \\ \hat{\theta} - \theta^* \end{bmatrix} \xrightarrow{D} \mathcal{N}\left(0, \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1} \Sigma_{1:T} \begin{bmatrix} \dot{\Phi}^*_{1:T-1} & 0 \\ V_{T,1:T-1} & \dot{\Psi}^* \end{bmatrix}^{-1,\top}\right). \blacksquare$$

C.4 Limit Theorems for Adaptively Sampled Data

Overview of Appendix C.4 Results.

- **Section C.4.1:** Weighted Martingale Triangular Array Law of Large Numbers (Theorem C.4.1)
- **Section C.4.2:** Weighted Martingale Triangular Array Uniform Law of Large Numbers (Theorem C.4.2)
- **Section C.4.3:** Showing Terms are $O_P(1)$ (Lemma C.4.1)
- **Section C.5:** Weighted Martingale Triangular Array Central Limit Theorem (Theorem C.5.1)
- **Section C.6:** Functional Asymptotic Normality under Finite Bracketing Integral (Theorem C.6.1)
- **Section C.6.1:** Stochastic Equicontinuity (Lemma C.6.1)

C.4.1 Weighted Martingale Triangular Array Law of Large Numbers (Theorem C.4.1)

Theorem C.4.1 (Weighted Martingale Triangular Array Weak Law of Large Numbers)**.** *Let f be a real-valued function of $\mathcal{H}_t^{(i)}$ such that for some $\alpha > 0$, $\mathbb{E}_{\pi^*_{2:t}}\left[\left|f\left(\mathcal{H}_t^{(i)}\right)\right|^{1+\alpha}\right] < \infty$. Under Condition 5.3.2 (Minimum Exploration),*

$$\frac{1}{n}\sum_{i=1}^{n} W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f(\mathcal{H}_t^{(i)}) \xrightarrow{P} \mathbb{E}_{\pi^*_{2:t}}\left[f(\mathcal{H}_t^{(i)})\right]. \tag{C.4.1}$$

Moreover,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) \xrightarrow{P} \mathbb{E}_{\pi_{2:t}^*}\left[\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]. \tag{C.4.2}$$

Proof of Theorem C.4.1.

Proving display (C.4.1) ***holds.*** To show display (C.4.1) holds it is sufficient to show that for any $t \in [1\colon T]$,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{ W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi_{2:t}^*}\left[f(\mathcal{H}_t^{(i)})\right]\right\} \xrightarrow{P} 0. \tag{C.4.3}$$

Above we use $W_1^{(i)} = 1$ and $W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \triangleq W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})$.

For the $t = 1$ case, $\mathcal{H}_1^{(1)}, \mathcal{H}_1^{(2)}, \mathcal{H}_1^{(3)}, \ldots, \mathcal{H}_1^{(n)}$ are i.i.d.; in this case, display (C.4.3) holds by the Weak Law of Large numbers for i.i.d. data.

Note that for $t \geq 2$, the distribution of $\mathcal{H}_t^{(i)}$ is changing with the number of users n, since it is depends on policy parameter $\hat{\beta}_{t-1}^{(n)}$. Thus, we need to consider triangular array asymptotics.

The first task is to rewrite the left-hand side of display (C.4.3) as a sum of triangular array martingale differences. Note that we can rewrite the left-hand side of display (C.4.3) as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{ W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f(\mathcal{H}_t^{(i)})\right]\right\}$$

Above the expectation $\mathbb{E}$ is with respect to the data distribution used to collect the data, thus, $\mathbb{E}\left[W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f(\mathcal{H}_t^{(i)})\right] = \mathbb{E}_{\pi_{2:t}^*}\left[f(\mathcal{H}_t^{(i)})\right]$.

Let $X_t^{(i)} \triangleq W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f(\mathcal{H}_t^{(i)})$. Also let $\mathcal{H}_0^{(1:n)} \triangleq \emptyset$ and $S_{T+1}^{(1:n)} \triangleq \emptyset$ (the second

definition is only used for the $t = T$ case). By telescoping series,

$$= \frac{1}{n} \sum_{i=1}^{n} \left\{ X_t^{(i)} - \mathbb{E}\left[X_t^{(i)} \right] \right\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left\{ \mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_0^{(1:n)}, S_1^{(1:n)} \right] - \mathbb{E}\left[X_t^{(i)} \right] \right\}}_{\triangleq Z_0^{(i)}}$$
$$+ \sum_{t'=1}^{t} \left[\frac{1}{n} \sum_{i=1}^{n} \underbrace{\left\{ \mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)} \right] - \mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)} \right] \right\}}_{\triangleq Z_{t'}^{(i)}} \right]. \quad \text{(C.4.4)}$$

Note above that $X_t^{(i)} = \mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)} \right]$, since $X_t^{(i)}$ is a constant given $\mathcal{H}_t^{(1:n)}$.

Using the $Z_{t'}^{(i)}$ notation defined in display (C.4.4) above,

$$= \underbrace{\frac{1}{n} \sum_{i=1}^{n} Z_0^{(i)}}_{= o_P(1)} + \frac{1}{n} \sum_{i=1}^{n} \sum_{t'=1}^{t} Z_{t'}^{(i)} \quad \text{(C.4.5)}$$

Note above that $\frac{1}{n}\sum_{i=1}^{n} Z_0^{(i)} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_0^{(1:n)}, S_1^{(1:n)} \right] - \mathbb{E}\left[X_t^{(i)} \right] \right\} =$
$\frac{1}{n}\sum_{i=1}^{n} \left\{ \mathbb{E}\left[f(\mathcal{H}_1^{(i)}) \middle| S_1^{(1:n)} \right] - \mathbb{E}\left[f(\mathcal{H}_1^{(i)}) \right] \right\} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \mathbb{E}\left[f(\mathcal{H}_1^{(i)}) \middle| S_1^{(i)} \right] - \mathbb{E}\left[f(\mathcal{H}_1^{(i)}) \right] \right\} \xrightarrow{P}$
0 by the weak law of large numbers for i.i.d. random variables.

Regarding the second summation in display (C.4.5), note that $\left\{ Z_{t'}^{(i)} \right\}_{i=1;t'=1}^{i=n;t'=t}$ is a martingale difference triangular array with respect to the filtration $\left\{ \sigma\left(\mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)} \right) \right\}_{t'=1}^{t}$. This is

the case because for any $i \in [1\colon n]$ and $t' \in [0\colon t]$,

$$\mathbb{E}\left[Z_{t'}^{(i)} \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right] = \mathbb{E}\left[\mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)}\right] - \mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right] \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right] \tag{C.4.6}$$

$$= \mathbb{E}\left[\mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)}\right] \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right] - \mathbb{E}\left[X_t^{(i)} \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right] = 0.$$

The final equality above holds by the law of iterated expectations.

The next step will be to apply Theorem 2(a) of[7] (Weak Law of Large Numbers for Triangular Array Mixingales). Theorem 2(a) ensures that $\frac{1}{n}\sum_{i=1}^{n}\sum_{t'=1}^{t} Z_{t'}^{(i)} \xrightarrow{P} 0$ if we can show the following hold:

(i) $\mathbb{E}\big[Z_{t'}^{(i)} \big| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\big] = 0$ for all $i \in [1\colon n]$ and $t' \in [1\colon t]$. (Note that since we have a martingale difference triangular array, the mixing constant $c_{n,i}$ in Theorem 2(a) is satisfied for $c_{n,i} = 0$.)

(ii) For some $\alpha > 0$, $\mathbb{E}\big[|Z_{t'}^{(i)}|^{1+\alpha}\big] < \infty$ for all $i \in [1\colon n]$ and $t' \in [1\colon t]$. Note that by Exercise 5.5.1 of[28] this implies $Z_{t'}^{(i)}$ are uniformly integrable.

We already showed that property (i) above holds earlier in display (C.4.6). All that remains is to show that property (ii) above holds.

Consider any $t' \in [1\colon t]$ and any $i \in [1\colon n]$,

$$\mathbb{E}\big[\big|Z_{t'}^{(i)}\big|^{1+\alpha}\big] = \mathbb{E}\left[\left|\mathbb{E}\big[X_t^{(i)} \big| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)}\big] - \mathbb{E}\big[X_t^{(i)} \big| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\big]\right|^{1+\alpha}\right]$$

By Lemma C.2.1, for some positive constant $c_{1+\alpha} < \infty$,

$$\leq c_{1+\alpha}\left\{\mathbb{E}\left[\left|\mathbb{E}\big[X_t^{(i)} \big| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)}\big]\right|^{1+\alpha}\right] + \mathbb{E}\left[\left|\mathbb{E}\big[X_t^{(i)} \big| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\big]\right|^{1+\alpha}\right]\right\}$$

By Jensen's inequality

$$\leq c_{1+\alpha}\left\{\mathbb{E}\left[\mathbb{E}\left[\left|X_t^{(i)}\right|^{1+\alpha}\middle|\mathcal{H}_{t'}^{(1:n)},S_{t'+1}^{(1:n)}\right]\right]+\mathbb{E}\left[\mathbb{E}\left[\left|X_t^{(i)}\right|^{1+\alpha}\middle|\mathcal{H}_{t'-1}^{(1:n)},S_{t'}^{(1:n)}\right]\right]\right\}$$

By the law of iterated expectations,

$$= c_{1+\alpha}\left\{\mathbb{E}\left[\left|X_t^{(i)}\right|^{1+\alpha}\right]+\mathbb{E}\left[\left|X_t^{(i)}\right|^{1+\alpha}\right]\right\} = c_{1+\alpha}2\mathbb{E}\left[\left|X_t^{(i)}\right|^{1+\alpha}\right]$$

Since $X_t^{(i)} \triangleq W_{1:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})f(\mathcal{H}_t^{(i)})$,

$$= c_{1+\alpha}2\mathbb{E}\left[\left|W_{1:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})f(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right]$$

Since $W_{1:t}^{(i)}(\beta^*,\hat{\beta}^{(n)}) \leq \pi_{\min}^{-(t-1)}$ by Condition 5.3.2 (Minimum Exploration),

$$\leq c_{1+\alpha}2\pi_{\min}^{-(t-1)(1+\alpha)}\mathbb{E}\left[\left|f(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right] < \infty.$$

The final inequality holds since $\mathbb{E}\left[\left|f(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right] < \infty$ by assumption of this Theorem. We have now shown that property (ii) holds. We have now shown that display (C.4.1) holds.

Proving Display (C.4.2) ***holds.***

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1}f(\mathcal{H}_t^{(i)}) \xrightarrow{P} \mathbb{E}_{\pi_{2:t}^*}\left[\left\{\pi_{2:t}^{*,(i)}\right\}^{-1}f(\mathcal{H}_t^{(i)})\right]. \tag{C.4.7}$$

Note that

$$\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1}f(\mathcal{H}_t^{(i)}) = W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})\left\{\pi_{2:t}^{*,(i)}\right\}^{-1}f(\mathcal{H}_t^{(i)}).$$

By Condition 5.3.2, $\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} \leq \pi_{\min}^{-(t-1)}$ a.s. Thus, $\mathbb{E}_{\pi_{2:t}^*}\left[\left|\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right] \leq \pi_{\min}^{-(t-1)}\mathbb{E}_{\pi_{2:t}^*}\left[\left|f(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right] < \infty$; the final inequality holds by the assumption that $\mathbb{E}_{\pi_{2:t}^*}\left[\left|f(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right] < \infty$. Thus, using the result from display (C.4.1),

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})$$
$$\xrightarrow{P} \mathbb{E}_{\pi_{2:t}^*}\left[\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]. \quad \blacksquare$$

C.4.2 Weighted Martingale Triangular Array Uniform Law of Large Numbers (Theorem C.4.2)

Theorem C.4.2 (Weighted Martingale Triangular Array Uniform Weak Law of Large Numbers)**.** *Let $\mathcal{F}$ be a class of real-valued, measurable functions such that for some $\alpha > 0$, $N_{[\,]}\left(\varepsilon, \mathcal{F}, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ for any $\varepsilon > 0$. Under Condition 5.3.2 (Minimum Exploration),*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi_{2:t}^*}\left[f(\mathcal{H}_t^{(i)})\right]\right\}\right| \xrightarrow{P} 0. \tag{C.4.8}$$

Moreover,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi_{2:t}^*}\left[\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]\right)\right| \xrightarrow{P} 0. \tag{C.4.9}$$

Proof of Theorem C.4.2.

Showing display (C.4.8) ***holds.*** Let $\varepsilon > 0$. Also let $N_\varepsilon \triangleq N_{[\,]}\left(\varepsilon, \mathcal{F}, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right)$. Since $N_{[\,]}\left(\varepsilon, \mathcal{F}, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ by assumption, we can finitely many brackets $\left\{(l_k, u_k)\right\}_{k=1}^{N_\varepsilon}$ that

cover $\mathcal{F}$ with $\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right]^{1/(1+\alpha)} \leq \varepsilon$. Thus,

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\left\{ W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})f(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi^*_{2:t}}\left[f(\mathcal{H}_t^{(i)})\right]\right\}$$

$$\leq \max_{k\in[1\colon N_\varepsilon]} \frac{1}{n}\sum_{i=1}^{n}\left\{ W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})u_k(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi^*_{2:t}}\left[l_k(\mathcal{H}_t^{(i)})\right]\right\}$$

By Theorem C.4.1 (Weighted Martingale Triangular Array Weak Law of Large Numbers), for any $k \in [1\colon N_\varepsilon]$, $\frac{1}{n}\sum_{i=1}^{n} W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})u_k(\mathcal{H}_t^{(i)}) \xrightarrow{P} \mathbb{E}_{\pi^*_{2:t}}\left[u_k(\mathcal{H}_t^{(i)})\right]$. Since there are finitely many brackets, by Slutsky's theorem this result holds simultaneously for all brackets $k \in [1\colon N_\varepsilon]$.

$$= o_P(1) + \max_{k\in[1\colon N_\varepsilon]}\left\{\mathbb{E}_{\pi^*_{2:t}}\left[u_k(\mathcal{H}_t^{(i)})\right] - \mathbb{E}_{\pi^*_{2:t}}\left[l_k(\mathcal{H}_t^{(i)})\right]\right\}$$

By Jensen's inequality

$$\leq o_P(1) + \max_{k\in[1\colon N_\varepsilon]} \mathbb{E}_{\pi^*_{2:t}}\left[|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})|^{1+\alpha}\right]^{1/(1+\alpha)} \leq o_P(1) + \varepsilon.$$

The last inequality above holds because our brackets were chosen to be at most of size ε, in $L_{1+\alpha}(\mathcal{P}_{\pi^*})$ norm. The above converges to zero because ε can be chosen to be arbitrarily small.

The final result display (C.4.8) holds by using the same argument above to show that $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi^*_{2:t}}\left[f(\mathcal{H}_t^{(i)})\right] - W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})f(\mathcal{H}_t^{(i)})\right\} \xrightarrow{P} 0$.

Showing display (C.4.9) ***holds.*** Note that

$$\begin{aligned}\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) &= W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) \\ &= W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \left\{ \prod_{t'=2}^{t} \frac{1}{\pi_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)}; \beta_{t'-1}^*)} \right\} f(\mathcal{H}_t^{(i)}).\end{aligned}$$

By the result from display (C.4.8), to show display (C.4.9) holds it is sufficient to show that the function class

$$\mathcal{F}_\pi \triangleq \left\{ \left(\prod_{t'=2}^{t} \frac{1}{\pi_{t'}(\,\cdot\,; \beta_{t'-1}^*)} \right) f(\,\cdot\,) \text{ s.t. } f \in \mathcal{F} \right\}$$

is such that $N_{[\,]}\left(\varepsilon, \mathcal{F}_\pi, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ for all $\varepsilon > 0$.

Since $N_\varepsilon \triangleq N_{[\,]}\left(\varepsilon, \mathcal{F}, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ by conditions of the theorem, we can find brackets $\left\{(l_k, u_k)\right\}_{k=1}^{N_\varepsilon}$ that cover $\mathcal{F}$ and $\mathbb{E}_{\pi_{2:t}^*}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right]^{1/(1+\alpha)} \leq \varepsilon$ for each bracket (l_k, u_k). Let $f \in \mathcal{F}$. We can find one of these brackets (l_k, u_k) such that $l_k(\mathcal{H}_t^{(i)}) \leq f(\mathcal{H}_t^{(i)}) \leq u_k(\mathcal{H}_t^{(i)})$ a.s.

Since $\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} > 0$ a.s., thus,

$$\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} l_k(\mathcal{H}_t^{(i)}) \leq \left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) \leq \left\{\pi_{2:t}^{*,(i)}\right\}^{-1} u_k(\mathcal{H}_t^{(i)}).$$

So the brackets $\left\{\left(\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} l_k, \left\{\pi_{2:t}^{*,(i)}\right\}^{-1} u_k\right)\right\}_{k=1}^{N_\varepsilon}$ cover $\mathcal{F}_\pi$.

Moreover, since $\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} \leq \pi_{\min}^{-(t-1)}$ a.s. by Condition 5.3.2, thus

$$\mathbb{E}_{\pi_{2:t}^*}\left[\left|\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} u_k(\mathcal{H}_t^{(i)}) - \left\{\pi_{2:t}^{*,(i)}\right\}^{-1} l_k(\mathcal{H}_t^{(i)})\right|^{1+\alpha}\right]^{1/(1+\alpha)} \leq \pi_{\min}^{-(t-1)} \varepsilon.$$

Thus, we have that $N_{[\,]}\left(\pi_{\min}^{-(t-1)}\varepsilon, \mathcal{F}_W, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) = N_{[\,]}\left(\varepsilon, \mathcal{F}, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right)$. Thus, since $N_{[\,]}\left(\varepsilon, \mathcal{F}, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ for any $\varepsilon > 0$ (by assumption of the Theorem), thus we have that $N_{[\,]}\left(\varepsilon, \mathcal{F}_W, L_{1+\alpha}(\mathcal{P}_{\pi^*})\right) < \infty$ for any $\varepsilon > 0$. ■

C.4.3 Showing Terms are $O_P(1)$ (Helper Lemma C.4.1)

Lemma C.4.1 (Showing Terms are $O_P(1)$ (Helper Lemma))**.** *Let g be a real-valued, function of $\mathcal{H}_t^{(i)}$ such that $\mathbb{E}_{\pi^*_{2:t}}\left[\left|g(\mathcal{H}_t^{(i)})\right|\right] < \infty$. Under Condition 5.3.2, we have that $g(\mathcal{H}_t^{(i)}) = O_P(1)$.*

Proof of Lemma C.4.1. Let $\varepsilon > 0$. By Condition 5.3.2 (Minimum Exploration),

$$W_{t'}^{(i)}\left(\beta^*_{t'-1}, \hat{\beta}^{(n)}_{t'-1}\right) = \frac{\pi^*_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)})}{\hat{\pi}^{(n)}_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)})} \geq \frac{\pi^*_{t'}(A_{t'}^{(i)}, S_{t'}^{(i)})}{1} \geq \pi_{\min} \quad \text{a.s.}$$

By the above result and Markov inequality, for any $c_\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|g(\mathcal{H}_t^{(i)})\right| > c_\varepsilon\right) &\leq c_\varepsilon^{-1}\mathbb{E}\left[\left|g(\mathcal{H}_t^{(i)})\right|\right] \\ &\leq c_\varepsilon^{-1}\pi_{\min}^{-(t-1)}\mathbb{E}\left[W_{2:t}^{(i)}\left(\beta^*, \hat{\beta}^{(n)}\right)\left|g(\mathcal{H}_t^{(i)})\right|\right] = c_\varepsilon^{-1}\pi_{\min}^{-(t-1)}\mathbb{E}_{\pi^*_{2:t}}\left[\left|g(\mathcal{H}_t^{(i)})\right|\right]. \quad \text{(C.4.10)}\end{aligned}$$

The above is less than or equal to ε by choosing $c_\varepsilon > \varepsilon^{-1}\pi_{\min}^{-(t-1)}\mathbb{E}_{\pi^*_{2:t}}\left[\left|g(\mathcal{H}_t^{(i)})\right|\right]$. This is sufficient by the definition of $O_P(1)$. ■

C.5 Weighted Martingale Triangular Array Central Limit Theorem (Theorem C.5.1)

Theorem C.5.1 (Weighted Martingale Triangular Array Central Limit Theorem)**.** *Let $f_1, f_2, \ldots, f_t$ be real-valued, measurable functions of $\mathcal{H}_1^{(i)}, \mathcal{H}_2^{(i)}, \ldots, \mathcal{H}_t^{(i)}$ respectively such that $\mathbb{E}_{\pi^*_{2:t'}}\left[\left|f_{t'}\left(\mathcal{H}_{t'}^{(i)}\right)\right|^{2+\alpha}\right] < \infty$ for some $\alpha > 0$, for each $t' \in [1\colon t]$. We show that (below we*

use $W_1^{(i)} = 1$*)*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \sum_{t'=1}^{t} W_{1:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_{t'}(\mathcal{H}_{t'}^{(i)}) - \mathbb{E}_{\pi^*_{2:t}}\left[\sum_{t'=1}^{t} f_{t'}(\mathcal{H}_{t'}^{(i)}) \right] \right\} \xrightarrow{D} \mathcal{N}\left(0, \mathrm{Var}_{\pi^*_{2:t}}\left(\sum_{t'=1}^{t} f_{t'}(\mathcal{H}_{t'}^{(i)}) \right)\right) \quad \text{(C.5.1)}$$

for $\mathrm{Var}_{\pi^*_{2:t}}\left(\sum_{t'=1}^{t} f_{t'}(\mathcal{H}_{t'}^{(i)})\right) \triangleq \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\sum_{t'=1}^{t} f_{t'}(\mathcal{H}_{t'}^{(i)})\right\}^2\right] - \mathbb{E}_{\pi^*_{2:t}}\left[\sum_{t'=1}^{t} f_{t'}(\mathcal{H}_{t'}^{(i)})\right]^2$ *under the following conditions:*

(A) Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) hold.

(B) $\hat{\beta}_{t'}^{(n)} - \beta_{t'}^* = O_P(1/\sqrt{n})$ *for all* $t' \in [1\colon t-1]$.

Proof of Theorem C.5.1. We want to show that display (C.5.1) holds for any $t \in [2\colon T]$. For notational convenience we consider the t set to T case; the argument holds by the same argument for any $t \in [2\colon T]$.

The first task is to rewrite the left-hand side of display (C.5.1) as a sum of triangular array martingale differences; we take an approach similar to that we used in the proof of Theorem C.4.1.

Note that the left-hand side of display (C.5.1) can be rewritten as follows:

$$\begin{aligned}\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \sum_{t'=1}^{t} W_{1:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_{t'}(\mathcal{H}_{t'}^{(i)}) - \mathbb{E}_{\pi^*_{2:t}}\left[\sum_{t'=1}^{t} f_{t'}(\mathcal{H}_{t'}^{(i)}) \right] \right\} \\ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) \right] \right\}\end{aligned}$$

Above we use $W_1^{(i)} = 1$ and $W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \triangleq W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})$. Additionally, above

the expectation $\mathbb{E}$ is with respect to the data distribution used to collect the data, thus, $\mathbb{E}\left[W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f(\mathcal{H}_t^{(i)})\right] = \mathbb{E}_{\pi_{2:t}^*}\left[f(\mathcal{H}_t^{(i)})\right]$.

Let $X_t^{(i)} \triangleq W_{1:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)})$ and $X_{1:T}^{(i)} \triangleq \sum_{t=1}^{T} X_t^{(i)}$.

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ X_t^{(i)} - \mathbb{E}\left[X_t^{(i)}\right] \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ X_{1:T}^{(i)} - \mathbb{E}\left[X_{1:T}^{(i)}\right] \right\}$$

Let $\mathcal{H}_0^{(1:n)} \triangleq \emptyset$ and $S_{T+1}^{(1:n)} \triangleq \emptyset$. Note that $X_{1:T}^{(i)} = \mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_T^{(1:n)}, S_{T+1}^{(1:n)}\right]$ since $X_{1:T}^{(i)}$ is known given $\mathcal{H}_T^{(1:n)}$. By telescoping series,

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\left\{ \mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_0^{(1:n)}, S_1^{(1:n)}\right] - \mathbb{E}\left[X_{1:T}^{(i)}\right] \right\}}_{\triangleq Z_0^{(i)}} + \sum_{t=1}^{T} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\left\{ \mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)}\right] - \mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] \right\}}_{\triangleq Z_t^{(i)}} \right].$$

Note that the terms $Z_t^{(i)}$ above are different from those we defined in the proof of Theorem C.4.1, since here the terms $X_{1:T}^{(i)}$ are a sum over T terms.

Note that $\mathbb{E}\left[Z_t^{(i)} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] = 0$ for all $i \in [1\colon n]$ and $t \in [1\colon T]$. This is the case because for any $i \in [1\colon n]$ and $t \in [1\colon T]$,

$$\mathbb{E}\left[Z_t^{(i)} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] = \mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)}\right] - \mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)}\right] \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] - \mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] = 0.$$

The final equality above holds by the law of iterated expectations.

In the next two subsections we will show the following two results:

(i) Convergence of conditional variance

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Z_0^{(i)}\right)^2\right]+\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Z_t^{(i)}\right)^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]\xrightarrow{P}\mathrm{Var}_{\pi^*_{2:T}}\left(\sum_{t=1}^{T}f_t(\mathcal{H}_t^{(i)})\right). \tag{C.5.2}$$

(ii) Conditional Lindeberg For any $\varepsilon>0$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Z_0^{(i)}\right)^2\mathbb{I}_{|Z_0^{(i)}|/\sqrt{n}>\varepsilon}\right]+\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Z_t^{(i)}\right)^2\mathbb{I}_{|Z_t^{(i)}|/\sqrt{n}>\varepsilon}\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]\xrightarrow{P}0. \tag{C.5.3}$$

With the above two results we can apply Theorem 2.2 of[29] (a martingale central limit theorem) to conclude that our desired result holds, i.e.,

$$\frac{1}{\sqrt{n}}\sum_{t=0}^{T}\sum_{i=1}^{n}Z_t^{(i)}\xrightarrow{D}\mathcal{N}\left(0,\mathrm{Var}_{\pi^*_{2:T}}\left(\sum_{t=1}^{T}f_t(\mathcal{H}_t^{(i)})\right)\right).$$

We first show that display (C.5.3) holds. We then show that display (C.5.2) holds.

(ii) Conditional Lindeberg; Display **(C.5.3).** For any $\varepsilon>0$, we show that the following conditional Lindeberg term is $o_P(1)$:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Z_0^{(i)}\right)^2\mathbb{I}_{|Z_0^{(i)}|/\sqrt{n}>\varepsilon}\right]+\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Z_t^{(i)}\right)^2\mathbb{I}_{|Z_t^{(i)}|/\sqrt{n}>\varepsilon}\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]$$

Note that for $\alpha>0$, $\mathbb{I}_{|Z|/\sqrt{n}>\varepsilon}=\mathbb{I}_{|Z|/(\varepsilon\sqrt{n})>1}=\mathbb{I}_{|Z|^\alpha/(\varepsilon\sqrt{n})^\alpha>1}\leq|Z|^\alpha/(\varepsilon\sqrt{n})^\alpha$.

Thus we can upper-bound the previous display as follows:

$$\leq \frac{1}{n(\varepsilon\sqrt{n})^{\alpha}} \sum_{i=1}^{n} \mathbb{E}\left[\left|Z_0^{(i)}\right|^{2+\alpha}\right] + \sum_{t=1}^{T} \frac{1}{n(\varepsilon\sqrt{n})^{\alpha}} \sum_{i=1}^{n} \mathbb{E}\left[\left|Z_t^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right]. \quad \text{(C.5.4)}$$

Note that for any $t' \in [0\colon T]$,

$$\left|Z_{t'}^{(i)}\right|^{2+\alpha} = \left|\mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)}\right] - \mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right]\right|^{2+\alpha}$$

Above, by slight abuse of notation, if $t' = 0$, we use $\mathcal{H}_{-1}^{(1:n)} \triangleq \emptyset$ and $S_0^{(1:n)} \triangleq \emptyset$.

$$= \left|\sum_{s=1}^{T} \left\{\mathbb{E}\left[X_s^{(i)} \middle| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)}\right] - \mathbb{E}\left[X_s^{(i)} \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right]\right\}\right|^{2+\alpha}$$

By repeatedly applying Lemma C.2.1 (Inequality Using Binomial Theorem) for any numbers $a_1, a_2, \ldots, a_K$, $|\sum_{k=1}^{K} a_k|^{\eta} \leq c_{2+\alpha}^{K} \sum_{k=1}^{K} |a_k|^{2+\alpha}$ for some constant $c_{2+\alpha} > 0$.

$$\leq c_{2+\alpha}^{T} \sum_{s=1}^{T} \left\{\left|\mathbb{E}\left[X_s^{(i)} \middle| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)}\right]\right|^{2+\alpha} + \left|\mathbb{E}\left[X_s^{(i)} \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right]\right|^{2+\alpha}\right\}$$

By Jensen's Inequality,

$$\leq c_{2+\alpha}^{T} \sum_{s=1}^{T} \left\{\mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t'}^{(1:n)}, S_{t'+1}^{(1:n)}\right] + \mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t'-1}^{(1:n)}, S_{t'}^{(1:n)}\right]\right\}$$

Thus, we can upper-bound display (C.5.4) as follows:

$$\frac{1}{n(\varepsilon\sqrt{n})^{\alpha}} \sum_{i=1}^{n} \mathbb{E}\left[\left|Z_0^{(i)}\right|^{2+\alpha}\right] + \sum_{t=1}^{T} \frac{1}{n(\varepsilon\sqrt{n})^{\alpha}} \sum_{i=1}^{n} \mathbb{E}\left[\left|Z_t^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right]$$

$$\leq \frac{c_{2+\alpha}^T}{n(\varepsilon\sqrt{n})^\alpha} \sum_{i=1}^n \sum_{s=1}^T \left\{ \mathbb{E}\left[\mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_0^{(1:n)}, S_1^{(1:n)} \right] + \mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \right] \right] \right\}$$
$$+ \sum_{t=1}^T \frac{c_{2+\alpha}^T}{n(\varepsilon\sqrt{n})^\alpha} \sum_{i=1}^n \sum_{s=1}^T \left\{ \mathbb{E}\left[\mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)} \right] \right.\right.$$
$$\left.\left. + \mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] \right\}$$

By the law of itereated expectations,

$$= \frac{c_{2+\alpha}^T}{n(\varepsilon\sqrt{n})^\alpha} \sum_{i=1}^n \sum_{s=1}^T 2\mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \right] + \sum_{t=1}^T \frac{c_{2+\alpha}^T}{n(\varepsilon\sqrt{n})^\alpha} \sum_{i=1}^n \sum_{s=1}^T 2\mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right]$$

To show that the above is $o_P(1)$, it is sufficient to show that $\mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right]$ and $\mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \right]$ for all $t, s \in [1\colon T]$ are all $O_P(1)$. By Lemma C.4.1, it is sufficient to show that $\mathbb{E}_{\pi_{2:s}^*}\left[\left|X_s^{(i)}\right|^{2+\alpha} \right]$ and $\mathbb{E}_{\pi_{2:t}^*}\left[\left| \mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] \right| \right]$ are bounded.

By Jensen's inequality,

$$\mathbb{E}_{\pi_{2:t}^*}\left[\left| \mathbb{E}\left[\left|X_s^{(i)}\right|^{2+\alpha} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] \right| \right] \leq \mathbb{E}_{\pi_{2:t}^*}\left[\left|X_s^{(i)}\right|^{2+\alpha} \right].$$

Thus, it is sufficient to show that $\mathbb{E}_{\pi_{2:t}^*}\left[\left|X_s^{(i)}\right|^{2+\alpha} \right] < \infty$ for all $s \in [1\colon T]$.

By Condition 5.3.2, $W_{2:s}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \leq \pi_{\min}^{-(s-1)}$ a.s. so,

$$\mathbb{E}_{\pi_{2:s}^*}\left[\left|X_s^{(i)}\right|^{2+\alpha} \right] = \mathbb{E}_{\pi_{2:s}^*}\left[\left| W_{2:s}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_s(\mathcal{H}_s^{(i)}) \right|^{2+\alpha} \right]$$

$$\leq \pi_{\min}^{-(s-1)(2+\alpha)} \mathbb{E}_{\pi_{2:s}^*}\left[W_{2:s}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \left| f_s(\mathcal{H}_s^{(i)}) \right|^{2+\alpha} \right]$$

$$= \pi_{\min}^{-(s-1)(2+\alpha)} \mathbb{E}_{\pi_{2:s}^*}\left[\left| f_s(\mathcal{H}_s^{(i)}) \right|^{2+\alpha} \right] < \infty.$$

(i) Convergence of conditional variance; Display (C.5.2)**:.** We now show that display (C.5.2) holds. Using the definition of $Z_t^{(i)}$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Z_0^{(i)}\right)^2\right]+\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Z_t^{(i)}\right)^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_0^{(1:n)},S_1^{(1:n)}\right]-\mathbb{E}\left[X_{1:T}^{(i)}\right]\right)^2\right]$$
$$+\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]-\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]\right)^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right].$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2\right]-\mathbb{E}\left[X_{1:T}^{(i)}\right]^2\right\}$$
$$+\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]-\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]^2\right\}. \tag{C.5.5}$$

Note by re-indexing, $-\sum_{t=1}^{T}\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]^2=-\sum_{t=0}^{T-1}\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]^2=$ $-\sum_{t=1}^{T}\underbrace{\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]^2}_{(a)}-\underbrace{\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2}_{(b)}+\underbrace{\left(X_{1:T}^{(i)}\right)^2}_{(c)}$. By rearranging terms the

terms in display (C.5.5),

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\underbrace{\left(X_{1:T}^{(i)}\right)^2}_{(c)} - \mathbb{E}\left[X_{1:T}^{(i)}\right]^2\Bigg\} \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2\right] - \underbrace{\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2}_{(b)}\right\} \\
&+ \sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)}\right]^2\middle|\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] - \underbrace{\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)}\right]^2}_{(a)}\right\}.
\end{aligned}
\tag{C.5.6}
$$

For the remainder of the proof we will show the following results, which combined with display (C.5.6) above are sufficient for display (C.5.2):

$$
\frac{1}{n}\sum_{i=1}^{n}\left\{\left(X_{1:T}^{(i)}\right)^2 - \mathbb{E}\left[X_{1:T}^{(i)}\right]^2\right\} \xrightarrow{P} \mathrm{Var}_{\pi^*_{2:T}}\left(\sum_{t=1}^{T} f_t(\mathcal{H}_t^{(i)})\right) \tag{C.5.7}
$$

$$
\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2\right] - \mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2\right\} \xrightarrow{P} 0 \tag{C.5.8}
$$

$$
\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)}\right]^2\middle|\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] - \mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)}\right]^2\right\} \xrightarrow{P} 0 \tag{C.5.9}
$$

Before showing the above three results, first note the following observations:

- By Condition 5.3.2 (Minimum Exploration),

$$\hat{\pi}_t^{(n)}(A_t^{(i)}, S_t^{(i)}) \geq \pi_{\min} \text{ a.s.} \quad \text{and} \quad W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \leq \pi_{\min}^{-(t-1)} \text{ a.s.} \tag{C.5.10}$$

- Additionally note that

$$\begin{aligned}
&\left| W_t^{(i)}(\beta_{t-1}^*, \hat{\beta}_{t-1}^{(n)}) - 1 \right| = \left| W_t^{(i)}(\beta_{t-1}^*, \hat{\beta}_{t-1}^{(n)}) - W_t^{(i)}(\beta_{t-1}^*, \beta_{t-1}^*) \right| \\
&\leq \left| \hat{\pi}_t^{(n)}\left(A_t^{(i)}, S_t^{(i)}\right)^{-1} - \pi_t^*\left(A_t^{(i)}, S_t^{(i)}\right)^{-1} \right| \leq \max_{a \in \mathcal{A}} \left| \hat{\pi}_t^{(n)}\left(a, S_t^{(i)}\right)^{-1} - \pi_t^*\left(a, S_t^{(i)}\right)^{-1} \right| \\
&\underbrace{\leq}_{(i)} \pi_{\min}^{-2} \max_{a \in \mathcal{A}} \left| \hat{\pi}_t^{(n)}\left(a, S_t^{(i)}\right) - \pi_t^*\left(a, S_t^{(i)}\right) \right| \\
&\underbrace{\leq}_{(ii)} \pi_{\min}^{-2} \max_{a \in \mathcal{A}} \dot{\pi}_t\left(a, S_t^{(i)}\right) \left\| \hat{\beta}_{t-1}^{(n)} - \beta_{t-1}^* \right\|_2.
\end{aligned} \tag{C.5.11}$$

  Inequality (i) above holds because by Taylor Series expansion, $\hat{\pi}^{-1} - \pi^{*,-1} = (-1)\tilde{\pi}^{-2}(\hat{\pi} - \pi^*)$ for some $\tilde{\pi}$ between $\hat{\pi}$ and π^*. By Condition 5.3.2 (Minimum Exploration) and display (C.5.10), $\tilde{\pi} \geq \min(\hat{\pi}, \pi^*) \geq \pi_{\min} > 0$ a.s.

  Inequality (ii) holds by Condition 5.3.3 (Lipschitz Policy Functions).

- Since the action space $\mathcal{A}$ is finite $\mathbb{E}_{\pi_{2:t}^*}\left[\max_{a \in \mathcal{A}} \dot{\pi}_t\left(a, S_t^{(i)}\right)\right] \leq \sum_{a \in \mathcal{A}} \mathbb{E}_{\pi_{2:t}^*}\left[\dot{\pi}_t\left(a, S_t^{(i)}\right)\right] < \infty$; the last inequality holds by Condition 5.3.3. Thus, by Lemma C.4.1, $\max_{a \in \mathcal{A}} \dot{\pi}_t\left(a, S_t^{(i)}\right) = O_P(1)$. Also since $\left\| \hat{\beta}_{t-1}^{(n)} - \beta_{t-1}^* \right\|_2 = O_P(1/\sqrt{n})$ by assumption, by display (C.5.11), we have that $W_t^{(i)}(\beta^*, \hat{\beta}^{(n)}) = 1 + O_P(1/\sqrt{n})$. Moreover, note that

$$W_{2:t}^{(i)}\left(\beta_{t-1}^*, \hat{\beta}_{t-1}^{(n)}\right) = \left(1 + O_P(1/\sqrt{n})\right)^{t-1} = 1 + O_P(1/\sqrt{n}). \tag{C.5.12}$$

1. Showing Display (C.5.7) *holds.* Since $X_{1:T}^{(i)} = \sum_{t=1}^{T} X_t^{(i)} = \sum_{t=1}^{T} W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)})$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_{1:T}^{(i)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{t=1}^{T} W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)})\right\}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T} W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)})\, W_{2:s}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_s(\mathcal{H}_s^{(i)})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T} W_{2:\max(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) f_s(\mathcal{H}_s^{(i)})\, W_{2:\min(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)})$$

Note that since $\mathbb{E}_{\pi_{2:t}^*}\left[f_t(\mathcal{H}_t^{(i)})^2\right]$ and $\mathbb{E}_{\pi_{2:s}^*}\left[f_s(\mathcal{H}_s^{(i)})^2\right]$ is bounded by assumption, by Lemma C.4.1, $f_t(\mathcal{H}_t^{(i)}) = O_P(1)$ and $f_s(\mathcal{H}_s^{(i)}) = O_P(1)$.

Moreover, by display (C.5.12), we have that $W_{2:\min(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) = 1 + O_P(1/\sqrt{n})$ and $W_{2:\max(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) = 1 + O_P(1/\sqrt{n})$.

Thus, $W_{2:\max(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) f_s(\mathcal{H}_s^{(i)})\, W_{2:\min(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)})$
$= W_{2:\max(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) f_s(\mathcal{H}_s^{(i)}) + O_P(1/\sqrt{n})$. So,

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{\sum_{s=1}^{T} W_{2:\max(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) f_s(\mathcal{H}_s^{(i)}) + O_P(1/\sqrt{n})\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T} W_{2:\max(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) f_s(\mathcal{H}_s^{(i)}) + o_P(1).$$

Note that by display (C.5.10),

$$\mathbb{E}_{\pi_{2:T}^*}\left[\left| W_{2:\max(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) f_s(\mathcal{H}_s^{(i)})\right|^{1+\alpha/2}\right]$$

$$\leq \pi_{\min}^{-\{\max(t,s)-1\}} \mathbb{E}_{\pi^*_{2:T}}\left[\left|f_t(\mathcal{H}_t^{(i)})f_s(\mathcal{H}_s^{(i)})\right|^{1+\alpha/2}\right]$$

$$\leq \pi_{\min}^{-\{\max(t,s)-1\}} \mathbb{E}_{\pi^*_{2:T}}\left[\max\left\{\left|f_t(\mathcal{H}_t^{(i)})\right|^2, \left|f_s(\mathcal{H}_s^{(i)})\right|^2\right\}^{1+\alpha/2}\right]$$

$$\leq \pi_{\min}^{-\{\max(t,s)-1\}} \mathbb{E}_{\pi^*_{2:T}}\left[\left\{\left|f_t(\mathcal{H}_t^{(i)})\right|^2 + \left|f_s(\mathcal{H}_s^{(i)})\right|^2\right\}^{1+\alpha/2}\right]$$

By Lemma C.2.1, for some positive constant $c_{1+\alpha/2}$,

$$\leq \pi_{\min}^{-\{\max(t,s)-1\}} c_{1+\alpha/2} \mathbb{E}_{\pi^*_{2:T}}\left[\left|f_t(\mathcal{H}_t^{(i)})\right|^{2+\alpha} + \left|f_s(\mathcal{H}_s^{(i)})\right|^{2+\alpha}\right] < \infty.$$

The above is bounded because of our assumption that $\mathbb{E}_{\pi^*_{2:T}}\left[\left|f_t(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right] < \infty$ and $\mathbb{E}_{\pi^*_{2:T}}\left[\left|f_s(\mathcal{H}_s^{(i)})\right|^{2+\alpha}\right] < \infty$. Thus, we can apply the Weighted Martingale Triangular Array Weak Law of Large Numbers (Theorem C.4.1) to get that

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_{1:T}^{(i)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T} W_{2:\max(t,s)}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_t(\mathcal{H}_t^{(i)}) f_s(\mathcal{H}_s^{(i)}) + o_P(1)$$
$$\xrightarrow{P} \sum_{t=1}^{T}\sum_{s=1}^{T} \mathbb{E}_{\pi^*_{2:T}}\left[f_t(\mathcal{H}_t^{(i)}) f_s(\mathcal{H}_s^{(i)})\right] = \mathbb{E}_{\pi^*_{2:T}}\left[\left\{\sum_{t=1}^{T} f_t(\mathcal{H}_t^{(i)})\right\}^2\right]$$

Additionally, note that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_{1:T}^{(i)}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi^*_{2:T}}\left[\sum_{t=1}^{T} f_t(\mathcal{H}_t^{(i)})\right]^2 = \mathbb{E}_{\pi^*_{2:T}}\left[\sum_{t=1}^{T} f_t(\mathcal{H}_t^{(i)})\right]^2.$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\left(X_{1:T}^{(i)}\right)^2-\mathbb{E}\left[X_{1:T}^{(i)}\right]^2\right\}\xrightarrow{P}\mathbb{E}_{\pi_{2:T}^*}\left[\left\{\sum_{t=1}^{T}f_t(\mathcal{H}_t^{(i)})\right\}^2\right]-\mathbb{E}_{\pi_{2:T}^*}\left[\sum_{t=1}^{T}f_t(\mathcal{H}_t^{(i)})\right]^2$$
$$=\mathrm{Var}_{\pi_{2:T}^*}\left(\sum_{t=1}^{T}f_t(\mathcal{H}_t^{(i)})\right).$$

The last equality above holds since $\mathrm{Var}_{\pi_{2:T}^*}\left(\sum_{t=1}^{T}f(\mathcal{H}_t^{(i)})\right)=\mathbb{E}_{\pi_{2:T}^*}\left[\left\{\sum_{t=1}^{T}f_t(\mathcal{H}_t^{(i)})\right\}^2\right]-\mathbb{E}_{\pi_{2:T}^*}\left[\sum_{t=1}^{T}f_t(\mathcal{H}_t^{(i)})\right]^2$.

2. Showing Display (C.5.8) *holds.*

Note that $\mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2\right]=\mathbb{E}\left[\mathbb{E}_{\pi_{2:T}^*}\left[\sum_{t=1}^{T}f_t(\mathcal{H}_t^{(i)})\middle|S_1^{(i)}\right]^2\right]$. Also, note the following:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2=\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{t=1}^{T}\mathbb{E}\left[W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})f_t(\mathcal{H}_t^{(i)})\middle|S_1^{(1:n)}\right]\right\}^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{t=1}^{T}\mathbb{E}_{\pi_{2:t}^*}\left[f_t(\mathcal{H}_t^{(i)})\middle|S_1^{(1:n)}\right]\right\}^2=\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{t=1}^{T}\mathbb{E}_{\pi_{2:t}^*}\left[f_t(\mathcal{H}_t^{(i)})\middle|S_1^{(i)}\right]\right\}^2$$

Since $\mathbb{E}_{\pi_{2:t}^*}\left[f_t(\mathcal{H}_t^{(i)})^2\right]<\infty$ by assumption, thus by Lemma C.2.1,

$\mathbb{E}_{\pi_{2:t}^*}\left[\left\{\sum_{t'=1}^{t}f_t(\mathcal{H}_t^{(i)})\right\}^2\right]<\infty$. Since $S_1^{(1)},S_1^{(2)},S_1^{(3)},\ldots S_1^{(n)}$ are i.i.d., by the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left\{\sum_{t=1}^{T}\mathbb{E}_{\pi_{2:t}^*}\left[f_t(\mathcal{H}_t^{(i)})\middle|S_1^{(i)}\right]\right\}^2-\mathbb{E}\left[\left\{\sum_{t=1}^{T}\mathbb{E}_{\pi_{2:T}^*}\left[f_t(\mathcal{H}_t^{(i)})\middle|S_1^{(i)}\right]\right\}^2\right]\right)\xrightarrow{P}0$$

Thus, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2\right]-\mathbb{E}\left[X_{1:T}^{(i)}\middle|S_1^{(1:n)}\right]^2\right\}\xrightarrow{P}0.$$

3. Showing Display (C.5.9) *holds.* Note that for any $t\in[1\colon T]$,

$$\mathbb{E}\left[X_{1:T}^{(i)}\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]^2=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{t'=1}^{t}X_{t'}^{(i)}+\sum_{t'=t+1}^{T}X_{t'}^{(i)}\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]^2$$

Recall $X_{t'}^{(i)}\triangleq W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})f_{t'}(\mathcal{H}_{t'}^{(i)})$.

$$=\left\{\sum_{t'=1}^{t}W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})f_{t'}(\mathcal{H}_{t'}^{(i)})+\mathbb{E}\left[\sum_{t'=t+1}^{T}W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]\right\}^2$$

$$=\left\{\sum_{t'=1}^{t}W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})f_{t'}(\mathcal{H}_{t'}^{(i)})+W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})\mathbb{E}\left[\sum_{t'=t+1}^{T}W_{t+1:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]\right\}^2$$

$$=\left\{\sum_{t'=1}^{t}W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})f_{t'}(\mathcal{H}_{t'}^{(i)})+W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})\mathbb{E}_{\pi^*_{t+1:T}}\left[\sum_{t'=t+1}^{T}f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]\right\}^2$$

Note $\mathbb{E}_{\pi^*_{t+1:T}}\left[\sum_{t'=t+1}^{T}f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(1:n)},S_{t+1}^{(1:n)}\right]=\mathbb{E}_{\pi^*_{t+1:T}}\left[\sum_{t'=t+1}^{T}f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(i)},S_{t+1}^{(i)}\right]$, since in the expectation actions are selected using fixed target policies $\pi^*_{t+1:T}$ (the first expec-

tation conditions on $\mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)}$ and the second conditions on $\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}$).

$$= \left\{ \sum_{t'=1}^{t} W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)}) f_{t'}(\mathcal{H}_{t'}^{(i)}) + W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \mathbb{E}_{\pi^*_{t+1:T}} \left[\sum_{t'=t+1}^{T} f_{t'}(\mathcal{H}_{t'}^{(i)}) \middle| \mathcal{H}_t^{(i)}, S_{t+1}^{(i)} \right] \right\}^2$$

For convenience, let $\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \triangleq f_{t'}(\mathcal{H}_{t'}^{(i)})$ for all $t' \in [1\colon t-1]$ and let $\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) \triangleq f_t(\mathcal{H}_t^{(i)}) + \mathbb{E}_{\pi^*_{t+1:T}} \left[\sum_{t'=t+1}^{T} f_{t'}(\mathcal{H}_{t'}^{(i)}) \middle| \mathcal{H}_t^{(i)}, S_{t+1}^{(i)} \right]$.

$$= \left\{ \sum_{t'=1}^{t} W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \right\}^2$$

$$= \underbrace{\left\{ \sum_{t'=1}^{t-1} W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \right\}^2}_{\triangleq (U_{1:t-1}^{(i)})^2} + \underbrace{\left\{ W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) \right\}^2}_{\triangleq (U_t^{(i)})^2}$$
$$+ \underbrace{2 \left\{ \sum_{t'=1}^{t-1} W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \right\} W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})}_{\triangleq U_{1:t-1}^{(i)} U_t^{(i)}} \quad \text{(C.5.13)}$$

Using the above result to rewrite $\mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)} \right]^2$, we can rewrite the left-hand side of display (C.5.9):

$$\sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}\left[\mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)} \right]^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] - \mathbb{E}\left[X_{1:T}^{(i)} \middle| \mathcal{H}_t^{(1:n)}, S_{t+1}^{(1:n)} \right]^2 \right\}$$

$$= \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \Bigg\{ \mathbb{E}\left[(U_{1:t-1}^{(i)})^2 + (U_t^{(i)})^2 + 2U_{1:t-1}U_t^{(i)} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] - (U_{1:t-1}^{(i)})^2 - (U_t^{(i)})^2 - 2U_{1:t-1}^{(i)}U_t^{(i)} \Bigg\}$$

Note that $(U_{1:t-1}^{(i)})^2$ is a constant given $\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}$ and cancel out in the display above. Thus,

$$= \frac{1}{n} 2 \sum_{i=1}^{n} \left\{ \mathbb{E}\left[U_{1:t-1}^{(i)} U_t^{(i)} \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] - U_{1:t-1}^{(i)} U_t^{(i)} \right\} + \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}\left[(U_t^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] - (U_t^{(i)})^2 \right\} \quad \text{(C.5.14)}$$

Before we show that display (C.5.14) converges in probability to zero, first note the following useful results:

- We first show that for all $t' \in [1\colon T]$,

$$\mathbb{E}_{\pi^*_{2:t'}} \left[\left| \tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \right|^{2+\alpha} \right] < \infty. \quad \text{(C.5.15)}$$

  For $t' \in [1\colon t-1]$, $\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) = f_{t'}(\mathcal{H}_{t'}^{(i)})$. Since $\mathbb{E}_{\pi^*_{2:t'}} \left[|f_{t'}(\mathcal{H}_{t'}^{(i)})|^{2+\alpha} \right] < \infty$ by assumption.

  For the $t' = t$ case, recall $\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) = f_t(\mathcal{H}_t^{(i)}) + \sum_{t'=t+1}^{T} \mathbb{E}_{\pi^*_{t+1:T}} \left[f_{t'}(\mathcal{H}_{t'}^{(i)}) \middle| \mathcal{H}_t^{(i)}, S_{t+1}^{(i)} \right]$. By repeatedly applying Lemma C.2.1 (Inequality Using Binomial Theorem) for any numbers $a_1, a_2, \ldots, a_K$, $|\sum_{k=1}^{K} a_k|^{2+\alpha} \leq c_{2+\alpha}^{K} \sum_{k=1}^{K} |a_k|^{2+\alpha}$ for some constant

$c_{2+\alpha} > 0$. Thus,

$$\begin{aligned}
&\mathbb{E}_{\pi^*_{2:t}}\left[\left|f_t(\mathcal{H}_t^{(i)}) + \sum_{t'=t+1}^{T} \mathbb{E}_{\pi^*_{t+1:T}}\left[f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\right]\right|^{2+\alpha}\right] \\
&\quad \le c_{2+\alpha}^{T-t}\mathbb{E}_{\pi^*_{2:t}}\left[\left|f_t(\mathcal{H}_t^{(i)})\right|^{2+\alpha} + \sum_{t'=t+1}^{T}\left|\mathbb{E}_{\pi^*_{t+1:T}}\left[f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\right]\right|^{2+\alpha}\right] \\
&\quad \le c_{2+\alpha}^{T-t}\sum_{t'=t}^{T}\mathbb{E}_{\pi^*_{2:T}}\left[\left|f_{t'}(\mathcal{H}_{t'}^{(i)})\right|^{2+\alpha}\right] < \infty.
\end{aligned}$$

The second to last inequality above holds by Jensen's inequality. The last inequality holds since $\mathbb{E}_{\pi^*_{2:t'}}\left[|f_{t'}(\mathcal{H}_{t'}^{(i)})|^{2+\alpha}\right] < \infty$ by assumption.

- We now show that for all $t', s \in [1\colon T]$,

$$\mathbb{E}_{\pi^*_{2:T}}\left[\left|\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)})\tilde{f}_s(\mathcal{H}_s^{(i)}, S_{s+1}^{(i)})\right|^{1+\alpha/2}\right] < \infty. \tag{C.5.16}$$

Note that for any real numbers a, b, that $ab \le \frac{1}{2}(a^2 + b^2)$. Thus,

$$\begin{aligned}
&\mathbb{E}_{\pi^*_{2:T}}\left[\left|\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)})\tilde{f}_s(\mathcal{H}_s^{(i)}, S_{s+1}^{(i)})\right|^{1+\alpha/2}\right] \\
&\qquad \le \frac{1}{2}\mathbb{E}_{\pi^*_{2:T}}\left[\left|\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)})\right|^{2+\alpha} + \left|\tilde{f}_s(\mathcal{H}_s^{(i)}, S_{s+1}^{(i)})\right|^{2+\alpha}\right] < \infty.
\end{aligned}$$

The last inequality holds by display (C.5.15).

We now show that $\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) = O_P(1)$.

- For $t' \in [1\colon t-1]$, $\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) = f_{t'}(\mathcal{H}_{t'}^{(i)})$ by definition. Since $\mathbb{E}_{\pi^*_{2:t'}}\left[|f_{t'}(\mathcal{H}_{t'}^{(i)})|\right] < \infty$ by assumption, by Lemma C.4.1, $\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) = f_{t'}(\mathcal{H}_{t'}^{(i)}) = O_P(1)$.

- Recall that $\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) = f_t(\mathcal{H}_t^{(i)}) + \sum_{t'=t+1}^{T} \mathbb{E}_{\pi^*_{t+1:T}}\left[f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\right]$ by definition. By Lemma C.4.1, $f_t(\mathcal{H}_t^{(i)}) = O_P(1)$ since $\mathbb{E}_{\pi^*_{2:t}}\left[|f_t(\mathcal{H}_t^{(i)})|\right] < \infty$ by assumption. Also note that by Jensen's inequality,

$$\mathbb{E}_{\pi^*_{2:t}}\left[\left|\mathbb{E}_{\pi^*_{t+1:T}}\left[f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\right]\right|\right] \leq \mathbb{E}_{\pi^*_{2:t}}\left[\mathbb{E}_{\pi^*_{t+1:T}}\left[|f_{t'}(\mathcal{H}_{t'}^{(i)})|\middle|\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\right]\right]$$
$$= \mathbb{E}_{\pi^*_{2:T}}\left[|f_{t'}(\mathcal{H}_{t'}^{(i)})|\right] < \infty.$$

Thus by Lemma C.4.1, we have that $\mathbb{E}_{\pi^*_{t+1:T}}\left[f_{t'}(\mathcal{H}_{t'}^{(i)})\middle|\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\right] = O_P(1)$. Combining the above results we have that $\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) = O_P(1)$.

3a. First Summation in Display (C.5.14)

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[U_{1:t-1}^{(i)}U_t^{(i)}\middle|\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] - U_{1:t-1}^{(i)}U_t^{(i)}\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[\left\{\sum_{t'=1}^{t-1} W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)})\right\} W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right]\right.$$
$$\left. - \left\{\sum_{t'=1}^{t-1} W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)})\right\} W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})\right)$$

$$= \sum_{t'=1}^{t-1}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)})\, W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right]\right.$$
$$\left. - W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)})\, W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})\right)$$

- Note the following for $t' \in [1\colon t-1]$:

$$\begin{aligned}
&\mathbb{E}\left[W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)},S_{t'+1}^{(i)})\,W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]\\
&\underbrace{=}_{(a)} W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})\,W_{2:t-1}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)},S_{t'+1}^{(i)})\mathbb{E}\left[W_t^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]\\
&\qquad\underbrace{=}_{(b)} W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})\,W_{2:t-1}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)},S_{t'+1}^{(i)})\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]\\
&\qquad\underbrace{=}_{(c)} W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})\,W_{2:t-1}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)},S_{t'+1}^{(i)})\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(i)},S_t^{(i)}\right] \quad \text{(C.5.17)}
\end{aligned}$$

- Equality (a) above holds since $t' < t$, so $W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})$ and $\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)},S_{t'+1}^{(i)})$ are constants given $\mathcal{H}_{t-1}^{(1:n)}$.

- Equality (b) above holds since the Radon-Nikodym weighting changes the policy with which actions are chosen with in the expectation.

- Regarding equality (c), note that in the expectation indexed by π_t^* that conditions on $\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}$, the only thing that is integrated over is the distribution of $(A_t^{(i)}, Y_t^{(i)})$. Given $\mathcal{H}_{t-1}^{(i)}, S_t^{(i)}$, when actions are selected using the policy π_t^*, the distribution of $(A_t^{(i)}, Y_t^{(i)})$ does not depend on the data of other users, i.e., $\mathcal{H}_{t-1}^{(j)}, S_t^{(j)}$ for $j \neq i$.

By display (C.5.17),

$$\begin{aligned}
= \sum_{t'=1}^{t-1}\frac{1}{n}\sum_{i=1}^{n}\Bigg(& W_{2:t-1}^{(i)}(\beta^*,\hat{\beta}^{(n)})\,W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)},S_{t'+1}^{(i)})\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(i)},S_t^{(i)}\right]\\
&- W_{2:t'}^{(i)}(\beta^*,\hat{\beta}^{(n)})\,W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)},S_{t'+1}^{(i)})\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})\Bigg)
\end{aligned}$$

By display (C.5.15), Jensen's inequality, and Lemma C.4.1, we have that $\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)},S_{t'+1}^{(i)})$

$= O_P(1), \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) = O_P(1)$ and $\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(i)}, S_t^{(i)}\right] = O_P(1)$. Additionally, by display (C.5.12), $W_{2:t'}^{(i)}(\beta^*, \hat{\beta}^{(n)}) = 1 + O_P(1/\sqrt{n})$. Thus,

$$= o_P(1) + \sum_{t'=1}^{t-1} \frac{1}{n} \sum_{i=1}^{n} \left(W_{2:t-1}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(i)}, S_t^{(i)}\right] \right.$$
$$\left. - W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) \right) \xrightarrow{P} 0.$$

The final limit above holds by the Weighted Martingale Triangular Array Weak Law of Large Numbers (Theorem C.4.1); note we can apply Theorem C.4.1 because we assume that Condition 5.3.2 holds and because $\mathbb{E}_{\pi_{2:t}^*}\left[\left|\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})\right|^{1+\alpha/2}\right] < \infty$ and

$\mathbb{E}_{\pi_{2:t}^*}\left[\left|\mathbb{E}_{\pi_t^*}\left[\tilde{f}_{t'}(\mathcal{H}_{t'}^{(i)}, S_{t'+1}^{(i)}) \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})\middle|\mathcal{H}_{t-1}^{(i)}, S_t^{(i)}\right]\right|^{1+\alpha/2}\right] < \infty$ by display (C.5.16) and Jensen's inequality.

3b. Second Summation in Display (C.5.14)

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}\left[(U_t^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] - (U_t^{(i)})^2 \right\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}\left[\left\{ W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) \right\}^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] \right.$$
$$\left. - \left\{ W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}) \right\}^2 \right)$$

Since $W_{2:t-1}^{(i)}(\beta^*, \hat{\beta}^{(n)})$ is a constant $\mathcal{H}_{t-1}^{(1:n)}$ and since by the Radon-Nikodym weighting, $\mathbb{E}\left[W_t^{(i)}(\beta^*, \hat{\beta}^{(n)})^2 \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right]$

$$= \mathbb{E}_{\pi_t^*}\left[W_t^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right],$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Bigg(W_{2:t-1}^{(i)}(\beta^*, \hat{\beta}^{(n)})^2 \mathbb{E}_{\pi_t^*}\left[W_t^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] - W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})^2 \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2\Bigg).$$

Note the following observations:

- By display (C.5.12), $W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)}) = 1 + O_P(1/\sqrt{n})$. By display (C.5.16) and Lemma C.4.1, we have that $\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 = O_P(1)$. Thus,

$$W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})^2 \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 = W_{2:t}^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 + O_P(1/\sqrt{n}). \tag{C.5.18}$$

- Additionally, for now, we take as given that

$$\mathbb{E}_{\pi_t^*}\left[W_t^{(i)}(\beta^*, \hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] = \mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(i)}, S_t^{(i)}\right] + O_P(1/\sqrt{n}). \tag{C.5.19}$$

  For now, we take as given that display (C.5.19) holds; we prove this at the end of this proof.

By displays (C.5.18) and (C.5.19) above,

$$= \frac{1}{n}\sum_{i=1}^{n}\Bigg(W_{2:t-1}^{(i)}(\beta^*,\hat{\beta}^{(n)})\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(i)},S_t^{(i)}\right] - W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2 + O_P(1/\sqrt{n})\Bigg)$$

$$= o_P(1) + \frac{1}{n}\sum_{i=1}^{n}\Bigg(W_{2:t-1}^{(i)}(\beta^*,\hat{\beta}^{(n)})\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(i)},S_t^{(i)}\right] - W_{2:t}^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\Bigg) \xrightarrow{P} 0.$$

The final limit above holds by the Weighted Martingale Triangular Array Weak Law of Large Numbers (Theorem C.4.1); note we can apply Theorem C.4.1 because we assume that Condition 5.3.2 holds and because by display (C.5.16) and Jensen's inequality, $\mathbb{E}_{\pi_{2:t}^*}\left[\left|\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\right|^{1+\alpha/2}\right] < \infty$ and $\mathbb{E}_{\pi_{2:t}^*}\left[\left|\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(i)},S_t^{(i)}\right]\right|^{1+\alpha/2}\right] < \infty$.

Now all that remains is to show that display (C.5.19) holds. We do this below:

- Note the following:

$$\begin{aligned}
&\mathbb{E}_{\pi_t^*}\left[W_t^{(i)}(\beta^*,\hat{\beta}^{(n)})\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right] \\
&= \mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right] + \mathbb{E}_{\pi_t^*}\left[\{W_t^{(i)}(\beta^*,\hat{\beta}^{(n)})-1\}\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right] \\
&= \mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(i)},S_t^{(i)}\right] + \mathbb{E}_{\pi_t^*}\left[\{W_t^{(i)}(\beta^*,\hat{\beta}^{(n)})-1\}\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]
\end{aligned} \tag{C.5.20}$$

  The first equality above holds by adding and subtracting $\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)},S_{t+1}^{(i)})^2\middle|\mathcal{H}_{t-1}^{(1:n)},S_t^{(1:n)}\right]$.

The second equality above holds because the expectation $\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right]$ only integrates over is the distribution of $(A_t^{(i)}, Y_t^{(i)})$; given $\mathcal{H}_{t-1}^{(i)}, S_t^{(i)}$, when actions are selected using the policy π_t^*, the distribution of $(A_t^{(i)}, Y_t^{(i)})$ does not depend on the data of other users, i.e., $\mathcal{H}_{t-1}^{(j)}, S_t^{(j)}$ for $j \neq i$.

- Now consider just the second term in the last line of display (C.5.20) above:

$$\begin{aligned}
&\mathbb{E}_{\pi_t^*}\left[\left\{W_t^{(i)}(\beta^*, \hat{\beta}^{(n)}) - 1\right\}\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] \\
&\qquad \leq \mathbb{E}_{\pi_t^*}\left[\left|W_t^{(i)}(\beta^*, \hat{\beta}^{(n)}) - 1\right| \tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] \\
&\underbrace{\leq}_{(a)} \pi_{\min}^{-2} \max_{a \in \mathcal{A}} \dot{\pi}_t(a, S_t^{(i)}) \mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] \left\|\hat{\beta}_{t-1}^{(n)} - \beta_{t-1}^*\right\|_2 \\
&\qquad\qquad \underbrace{=}_{(b)} O_P(1/\sqrt{n}).
\end{aligned}$$

Above inequality (a) holds by display (C.5.11); we are able to move the terms $\max_{a \in \mathcal{A}} \dot{\pi}_t(a, S_t^{(i)})$ and $\left\|\hat{\beta}_{t-1}^{(n)} - \beta_{t-1}^*\right\|_2$ out of the conditional expectation they are known given $\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}$.

Above limit (b) holds because (i) $\hat{\beta}_{t-1}^{(n)} = O_P(1/\sqrt{n})$ by assumption, (ii) $\mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] = \mathbb{E}_{\pi_t^*}\left[\tilde{f}_t(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)})^2 \middle| \mathcal{H}_{t-1}^{(i)}, S_t^{(i)}\right] = O_P(1)$ by display (C.5.16), Jensen's inequality, and Lemma C.4.1, and

(iii) $\max_{a\in\mathcal{A}} \dot{\pi}_t\big(a, S_t^{(i)}\big) = O_P(1)$; this holds by Lemma C.4.1 since

$$\begin{aligned}
&\mathbb{E}_{\pi^*_{2:t-1}}\left[\Big|\max_{a\in\mathcal{A}} \dot{\pi}_t\big(a, S_t^{(i)}\big)\Big|\right] \underbrace{\leq}_{(a)} \sum_{a\in\mathcal{A}} \mathbb{E}_{\pi^*_{2:t-1}}\left[\Big|\dot{\pi}_t\big(a, S_t^{(i)}\big)\Big|\right] \\
&= |\mathcal{A}| \sum_{a\in\mathcal{A}} \mathbb{E}_{\pi^*_{2:t-1}}\left[|\mathcal{A}|^{-1}\Big|\dot{\pi}_t\big(a, S_t^{(i)}\big)\Big|\right] \underbrace{=}_{(b)} |\mathcal{A}|\mathbb{E}_{\pi^*_{2:t-1},\pi_t^{\text{uniform}}}\left[\Big|\dot{\pi}_t\big(A_t^{(i)}, S_t^{(i)}\big)\Big|\right] \\
&\qquad \underbrace{=}_{(c)} |\mathcal{A}|\mathbb{E}_{\pi^*_{2:t}}\left[\frac{\pi_t^{\text{uniform}}(A_t^{(i)}, S_t^{(i)})}{\pi_t^*(A_t^{(i)}, S_t^{(i)})}\Big|\dot{\pi}_t\big(a, S_t^{(i)}\big)\Big|\right] \\
&\qquad\qquad \underbrace{\leq}_{(d)} |\mathcal{A}|\pi_{\min}^{-1}\mathbb{E}_{\pi^*_{2:t}}\left[\Big|\dot{\pi}_t\big(A_t^{(i)}, S_t^{(i)}\big)\Big|\right] \underbrace{\leq}_{(e)} \infty.
\end{aligned}$$

Inequality (a) above holds since the action space $\mathcal{A}$ is a finite set.

Equality (b) holds for $\pi_t^{\text{uniform}}(A_t^{(i)}, S_t^{(i)}) \triangleq |\mathcal{A}|^{-1}$, i.e., the policy that selects action uniformly over the action space $\mathcal{A}$ for the t^{th} action.

Equality (c) uses Radon-Nikodym derivative weights.

Inequality (d) above holds by exploration Condition 5.3.2.

Inequality (e) holds by Condition 5.3.3. ■

C.6 Functional Asymptotic Normality under Finite Bracketing Integral (Theorem C.6.1)

Theorem C.6.1 (Functional Asymptotic Normality under Finite Bracketing Integral for Adaptively Sampled Data)**.** *Let $\mathcal{F}$ be any class of real-valued measurable functions of $\mathcal{H}_t^{(i)}$ such that for all $f \in \mathcal{F}$, for some $\alpha > 0$, $\int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon, \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*})\big)}d\varepsilon < \infty$. Let Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) hold and*

also let $\hat{\beta}_{t'}^{(n)} - \beta_{t'}^* = O_P(1/\sqrt{n})$ *for all* $t' \in [1\colon t-1]$. *Then for*

$$\mathcal{G}_{\mathcal{F}}^{(n)}(f) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} f(\mathcal{H}_t^{(i)}) \right] \right),$$

the empirical process $\left\{ \mathcal{G}_{\mathcal{F}}^{(n)}(f) \,:\, f \in \mathcal{F} \right\}$ *converges in distribution to* $\mathcal{G}_{\mathcal{F}}$ *a mean-zero Gaussian process in* $\ell^\infty(\mathcal{F})$ *(the collection of all bounded functions from* $\mathcal{F}$ *to* $\mathbb{R}$*) with the following covariance function:*

$$\mathbb{E}\left[\mathcal{G}_{\mathcal{F}}(f) \mathcal{G}_{\mathcal{F}}(g) \right] \triangleq \mathbb{E}_{\pi_{2:t}^*} \Bigg[\left(\left\{ \pi_{2:t}^{*,(i)} \right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi_{2:t}^*} \left[\left\{ \pi_{2:t}^{*,(i)} \right\}^{-1} f(\mathcal{H}_t^{(i)}) \right] \right) \left(\left\{ \pi_{2:t}^{*,(i)} \right\}^{-1} g(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi_{2:t}^*} \left[\left\{ \pi_{2:t}^{*,(i)} \right\}^{-1} g(\mathcal{H}_t^{(i)}) \right] \right) \Bigg]. \quad \text{(C.6.1)}$$

Above we use $\pi_{2:t}^{*,(i)} \triangleq \prod_{t'=2}^{t} \pi_{t'}^*(A_{t'}^{(i)}, S_{t'}^{(i)})$.

Proof of Theorem C.6.1. By[99] Theorem 18.14, to show the desired result it is sufficient to show that the following two properties hold:

(a) **Joint Convergence of Marginals** For any finite number of functions $f_1, f_2, \ldots, f_K \in \mathcal{F}$,

$$\left(\mathcal{G}_{\mathcal{F}}^{(n)}(f_1), \mathcal{G}_{\mathcal{F}}^{(n)}(f_2), \ldots, \mathcal{G}_{\mathcal{F}}^{(n)}(f_K) \right) \xrightarrow{D} \left(\mathcal{G}_{\mathcal{F}}(f_1), \mathcal{G}_{\mathcal{F}}(f_2), \ldots, \mathcal{G}_{\mathcal{F}}(f_K) \right)$$

(b) **Asymptotically Tight** For any $\varepsilon, \eta > 0$, there exists a partition of $\mathcal{F}$ into finitely many sets $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_J$ such that

$$\limsup_{n \to \infty} \mathbb{P}^* \left(\sup_{j \in [1\colon J]} \sup_{f, f' \in \mathcal{F}_j} \left| \mathcal{G}_{\mathcal{F}}^{(n)}(f) - \mathcal{G}_{\mathcal{F}}^{(n)}(f') \right| > \varepsilon \right) \leq \eta.$$

Showing (a) Joint Convergence of Marginals. We can show that (a) above holds for the stochastic process $\left\{\mathcal{G}_{\mathcal{F}}^{(n)}(f) \,:\, f \in \mathcal{F}\right\}$ by the Theorem C.5.1 (Weighted Martingale Triangular Array Central Limit Theorem).

Specifically, by Cramer Wold device, it is sufficient to show that for any $c = [c_1, c_2, \ldots, c_K] \in \mathbb{R}^K$ that

$$\sum_{k=1}^{K} c_k \mathcal{G}_{\mathcal{F}}^{(n)}(f_k) \xrightarrow{D} \mathcal{N}\left(0, c^\top \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \ldots & \Sigma_{1,K} \\ \Sigma_{2,1} & \Sigma_{2,2} & \ldots & \Sigma_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{K,1} & \Sigma_{K,2} & \ldots & \Sigma_{K,K} \end{bmatrix} c\right)$$

where $\Sigma_{k,k'} \triangleq \mathbb{E}_{\pi^*_{2:t}}\left[\mathcal{G}_{\mathcal{F}}(f_k)\mathcal{G}_{\mathcal{F}}(f_{k'})\right]$.

Note that

$$\sum_{k=1}^{K} c_k \mathcal{G}_{\mathcal{F}}^{(n)}(f_k) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} \sum_{k=1}^{K} c_k f_k(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} \sum_{k=1}^{K} c_k f_k(\mathcal{H}_t^{(i)}) \right] \right)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(W_{2:t}^{(i)}\left(\beta^*, \hat{\beta}^{(n)}\right) \left\{\pi_{2:t}^{*,(i)}\right\}^{-1} \sum_{k=1}^{K} c_k f_k(\mathcal{H}_t^{(i)}) \right.$$
$$\left. - \mathbb{E}\left[W_{2:t}^{(i)}\left(\beta^*, \hat{\beta}^{(n)}\right) \left\{\pi_{2:t}^{*,(i)}\right\}^{-1} \sum_{k=1}^{K} c_k f_k(\mathcal{H}_t^{(i)}) \right] \right) \xrightarrow{D} \mathcal{N}\left(0, \bar{\Sigma}\right),$$

where

$$\bar{\Sigma} \triangleq \mathbb{E}_{\pi^*_{2:t}}\left[\left(\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} \sum_{k=1}^{K} c_k f_k(\mathcal{H}_t^{(i)}) \right)^2 \right] - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} \sum_{k=1}^{K} c_k f_k(\mathcal{H}_t^{(i)}) \right]^2. \quad \text{(C.6.2)}$$

The above weak convergence holds by Theorem C.5.1 (Weighted Martingale Triangular Array Central Limit Theorem). When applying Theorem C.5.1, we use the following properties:

- Conditions 5.3.2 (Minimum Exploration) and 5.3.3 (Lipschitz Policy Functions) hold.
- $\hat{\beta}_{t'}^{(n)} - \beta_{t'}^{*} = O_P(1/\sqrt{n})$ for all $t' \in [1\colon t-1]$.
- Recall that for some $\alpha > 0$, $\int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon, \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*})\big)}\, d\varepsilon < \infty$. This means that for all $f \in \mathcal{F}$, we can find a bracket (l_j, u_j) such that $l_j(\mathcal{H}_t^{(i)}) \leq f(\mathcal{H}_t^{(i)}) \leq u_j(\mathcal{H}_t^{(i)})$ a.s. Additionally, recall that the brackets are such that $\mathbb{E}_{\pi_{2:t}^*}\big[\big|l\big(\mathcal{H}_t^{(i)}\big)\big|^{2+\alpha}\big] < \infty$ and $\mathbb{E}_{\pi_{2:t}^*}\big[\big|u\big(\mathcal{H}_t^{(i)}\big)\big|^{2+\alpha}\big] < \infty$. Thus, for all $f \in \mathcal{F}$,

$$\mathbb{E}_{\pi_{2:t}^*}\big[\big|f\big(\mathcal{H}_t^{(i)}\big)\big|^{2+\alpha}\big] \leq \mathbb{E}_{\pi_{2:t}^*}\big[\big|l\big(\mathcal{H}_t^{(i)}\big)\big|^{2+\alpha}\big] + \mathbb{E}_{\pi_{2:t}^*}\big[\big|u\big(\mathcal{H}_t^{(i)}\big)\big|^{2+\alpha}\big] < \infty.$$

  Thus, by repeatedly applying Lemma C.2.1, for some positive constant $c_{2+\alpha} < \infty$, the following result holds:

$$\begin{aligned}\mathbb{E}_{\pi_{2:t}^*}\left[\left|\left\{\pi_{2:t}^{*,(i)}\right\}^{-1}\sum_{k=1}^{K} c_k f_k(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right] &\leq c_{2+\alpha}^{K}\sum_{k=1}^{K}\mathbb{E}_{\pi_{2:t}^*}\left[\left|\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} c_k f_k(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right] \\ &\leq c_{2+\alpha}^{K}\sum_{k=1}^{K}\pi_{\min}^{(t-1)(2+\alpha)}|c_k|^{2+\alpha}\mathbb{E}_{\pi_{2:t}^*}\left[\left|f_k(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right] < \infty.\end{aligned}$$

  The last inequality above holds by Condition 5.3.2.

By the definition of $\bar{\Sigma}$ from display (C.6.2),

$$\bar{\Sigma} = \mathbb{E}_{\pi^*_{2:t}}\left[\left(\left\{\pi^{*,(i)}_{2:t}\right\}^{-1}\sum_{k=1}^{K} c_k f_k(\mathcal{H}^{(i)}_t)\right)^2\right] - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1}\sum_{k=1}^{K} c_k f_k(\mathcal{H}^{(i)}_t)\right]^2$$

$$= \sum_{k=1}^{K}\sum_{k'=1}^{K}\left\{\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} c_k f_k(\mathcal{H}^{(i)}_t)\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} c_{k'} f_{k'}(\mathcal{H}^{(i)}_t)\right]\right.$$
$$\left. - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} c_k f_k(\mathcal{H}^{(i)}_t)\right]\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} c_{k'} f_{k'}(\mathcal{H}^{(i)}_t)\right]\right\}$$

$$= \sum_{k=1}^{K}\sum_{k'=1}^{K} c_k c_{k'}\mathbb{E}_{\pi^*_{2:t}}\left[\left(\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f_k(\mathcal{H}^{(i)}_t) - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f_k(\mathcal{H}^{(i)}_t)\right]\right)\right.$$
$$\left.\left(\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f_{k'}(\mathcal{H}^{(i)}_t) - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f_{k'}(\mathcal{H}^{(i)}_t)\right]\right)\right]$$

By the definition of $\mathbb{E}\left[\mathcal{G}_{\mathcal{F}}(f_k)\mathcal{G}_{\mathcal{F}}(f_{k'})\right]$ from display (C.6.1),

$$= \sum_{k=1}^{K}\sum_{k'=1}^{K} c_k c_{k'}\mathbb{E}\left[\mathcal{G}_{\mathcal{F}}(f_k)\mathcal{G}_{\mathcal{F}}(f_{k'})\right] = \sum_{k=1}^{K}\sum_{k'=1}^{K} c_k c_{k'}\Sigma_{k,k'} = c^\top \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \dots & \Sigma_{1,K} \\ \Sigma_{2,1} & \Sigma_{2,2} & \dots & \Sigma_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{K,1} & \Sigma_{K,2} & \dots & \Sigma_{K,K} \end{bmatrix} c.$$

The second equality above holds since recall $\Sigma_{k,k'} \triangleq \mathbb{E}\left[\mathcal{G}_{\mathcal{F}}(f_k)\mathcal{G}_{\mathcal{F}}(f_{k'})\right]$.

Thus, we have shown that $\bar{\Sigma}$ from display (C.6.2) is such that

$$\bar{\Sigma} = c^\top \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \dots & \Sigma_{1,K} \\ \Sigma_{2,1} & \Sigma_{2,2} & \dots & \Sigma_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{K,1} & \Sigma_{K,2} & \dots & \Sigma_{K,K} \end{bmatrix} c.$$

Showing (b) Asymptotically Tight. The asymptotically tight condition above holds by the same argument used in the proof of Theorem 19.5 from[99], but by replacing the use of maximal inequality Lemma 19.34 of[99] in that proof with our maximal inequality from Lemma C.8.1 (Maximal Inequality as a Function of the Bracketing Integral). We discuss this argument below.

Let $\varepsilon, \delta > 0$. Since $\mathcal{F}$ has finite bracketing integral by assumption, we can find a partition of $\mathcal{F}$ into finitely many sets $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_{N_{\delta_0}}$ where $N_{\delta_0} \triangleq N_{[\,]}\big(\delta_0, \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*})\big)$ and $\delta_0 \triangleq \delta\sqrt{\pi_{\min}^{-(t-1)}}$. Note that

$$\mathbb{P}^*\left(\sup_{j\in[1\colon N_{\delta_0}]} \sup_{f,f'\in\mathcal{F}_j} \left| \mathcal{G}_{\mathcal{F}}^{(n)}(f) - \mathcal{G}_{\mathcal{F}}^{(n)}(f') \right| > \varepsilon \right)$$

Above $\mathbb{P}^*$ refers to outer probabilities as defined in Section 18.2[99].

$$= \mathbb{P}^*\left(\sup_{j\in[1\colon N_{\delta_0}]} \sup_{f,f'\in\mathcal{F}_j} \left| \mathcal{G}_{\mathcal{F}}^{(n)}(f - f') \right| > \varepsilon \right)$$

By Markov inequality,

$$\leq \frac{1}{\varepsilon}\mathbb{E}^*\left[\sup_{j\in[1:\,N_{\delta_0}]}\sup_{f,f'\in\mathcal{F}_j}\left|\mathcal{G}_{\mathcal{F}}^{(n)}(f-f')\right|\right].$$

Let $\mathcal{G}_\delta$ be the function class such that $\mathcal{G}_\delta \triangleq \left\{f-f' \text{ s.t. } f,f' \in \mathcal{F}_j, j\in[1:\,N_{\delta_0}]\right\}$. Note the following observations:

- Note that for any $f,f' \in \mathcal{F}_j$, $\mathbb{E}_{\pi^*_{2:t}}\left[\left\{f(\mathcal{H}_t^{(i)}) - f'(\mathcal{H}_t^{(i)})\right\}^2\right] \leq \delta_0^2$; thus $\mathbb{E}_{\pi^*_{2:t}}\left[g(\mathcal{H}_t^{(i)})^2\right] \leq \delta_0^2$ for any $g\in\mathcal{G}_\delta$. Note that by Condition 5.3.2, this implies that $\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} g(\mathcal{H}_t^{(i)})^2\right] \leq \pi_{\min}^{-(t-1)}\mathbb{E}_{\pi^*_{2:t}}\left[g(\mathcal{H}_t^{(i)})^2\right] \leq \pi_{\min}^{-(t-1)}\delta_0^2 = \delta^2$ for all $g\in\mathcal{G}_\delta$
- We take as given for now that

$$\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon, \mathcal{G}_\delta, L_2\left(\mathcal{P}_{\pi^*}\right)\right)}d\varepsilon < \infty. \tag{C.6.3}$$

  We show the above holds at the end of this proof.

- We also take as given that there exists a non-negative envelope function G where $\left|g(\mathcal{H}_t^{(i)})\right| \leq G(\mathcal{H}_t^{(i)})$ a.s. for all $g\in\mathcal{G}_\delta$ and $\mathbb{E}_{\pi^*_{2:t}}\left[G(\mathcal{H}_t^{(i)})^2\right] < \infty$ (we show this at the end of this proof).

Using the above observations and Condition 5.3.2, we can apply Lemma C.8.1 (Maximal Inequality as a Function of the Bracketing Integral) to get that

$$\lesssim \frac{1}{\varepsilon}\left\{\int_0^\delta \sqrt{\log N_{[\,]}\left(\eta, \mathcal{G}_\delta, L_2\left(\mathcal{P}_{\pi^*}\right)\right)}d\eta + \sqrt{n}\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} G(\mathcal{H}_t^{(i)})\mathbb{I}_{G(\mathcal{H}_t^{(i)})>\sqrt{n}a(\delta)}\right]\right\}, \tag{C.6.4}$$

where $a(\delta) \triangleq \delta/\sqrt{\log N_{[]}\big(\delta, \mathcal{G}_\delta, L_2(\mathcal{P}_{\pi^*})\big)}$. Above $\lesssim$ means less than or equal to when scaled by universal positive constants.

- By display (C.6.3), the first term (integral term) in display (C.6.4) converges to zero as $\delta \to 0$.

- Regarding the second term in display (C.6.4), since $\mathbb{I}_{G(\mathcal{H}_t^{(i)})>\sqrt{n}a(\delta)} = 1$ implies that $G(\mathcal{H}_t^{(i)})\big\{\sqrt{n}a(\delta)\big\}^{-1} > 1$, thus, $\big|G(\mathcal{H}_t^{(i)})\big\{\sqrt{n}a(\delta)\big\}^{-1}\big| \geq \mathbb{I}_{G(\mathcal{H}_t^{(i)})>\sqrt{n}a(\delta)}$.

$$\sqrt{n}\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} G(\mathcal{H}_t^{(i)})\mathbb{I}_{G(\mathcal{H}_t^{(i)})>\sqrt{n}a(\delta)}\right] \leq a(\delta)^{-1}\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} G(\mathcal{H}_t^{(i)})^2\right]$$
$$\leq a(\delta)^{-1}\pi_{\min}^{-(t-1)}\mathbb{E}_{\pi^*_{2:t}}\left[G(\mathcal{H}_t^{(i)})^2\right].$$

  The last inequality above holds by Condition 5.3.2. The above goes to zero as $n \to \infty$ for every fixed δ.

Thus, we have that display (C.6.4) converges to zero; this is sufficient for the Theorem to hold. All that remains is to show that display (C.6.3) holds and that we can find an envelope function G. We do this below.

Bracketing Functions for $\mathcal{G}_\delta$; display (C.6.3). Let $\eta > 0$. Since $\mathcal{F}$ has finite bracketing integral by assumption, we can find $N_\eta \triangleq N_{[]}\big(\eta, \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*})\big)$ bracketing functions $\big\{(l_k, u_k)\big\}_{k=1}^{N_\eta}$. Note that for any $g \in \mathcal{G}_\delta$, we can find some $f, f' \in \mathcal{F}$ such that $g = f - f'$.

We will show that the brackets $\big\{(l_k - u_{k'},\ u_k - l_{k'})\big\}_{k=1;k'=1}^{k=N_\eta;k'=N_\eta}$ will cover $\mathcal{G}_\delta$ and be of size $2\sqrt{c_2}\eta$ in $L_{2+\alpha}(\mathcal{P}_{\pi^*})$ norm for a positive constant $c_2 < \infty$.

Covering: We can find brackets (l_k, u_k) and $(l_{k'}, u_{k'})$ such that $l_k(\mathcal{H}_t^{(i)}) \leq f(\mathcal{H}_t^{(i)}) \leq u_k(\mathcal{H}_t^{(i)})$ a.s. and $l_{k'}(\mathcal{H}_t^{(i)}) \leq f'(\mathcal{H}_t^{(i)}) \leq u_{k'}(\mathcal{H}_t^{(i)})$ a.s. Thus, $l_k(\mathcal{H}_t^{(i)}) - u_{k'}(\mathcal{H}_t^{(i)}) \leq$

$f(\mathcal{H}_t^{(i)}) - f'(\mathcal{H}_t^{(i)}) \leq u_k(\mathcal{H}_t^{(i)}) - l_{k'}(\mathcal{H}_t^{(i)})$ a.s.

Size: Note that

$$\sqrt{\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_{k'}(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)}) + u_{k'}(\mathcal{H}_t^{(i)})\right|^2\right]}$$

By Lemma C.2.1 for some positive constant $c_2 < \infty$,

$$\leq \sqrt{c_2\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^2\right] + c_2\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_{k'}(\mathcal{H}_t^{(i)}) - l_{k'}(\mathcal{H}_t^{(i)})\right|^2\right]}$$

Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$ (to see this square both sides),

$$\leq \sqrt{c_2\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^2\right]} + \sqrt{c_2\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_{k'}(\mathcal{H}_t^{(i)}) - l_{k'}(\mathcal{H}_t^{(i)})\right|^2\right]}$$

$$\leq 2\sqrt{c_2}\eta.$$

We now discuss why the last inequality above holds. By construction of our bracketing functions, $\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right]^{1/(2+\alpha)} \leq \eta$. Note that $h(x) = x^{1+\alpha/2}$ is convex. By Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^2\right]^{1+\alpha/2} &= h\left(\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^2\right]\right) \\ &\leq \mathbb{E}_{\pi^*_{2:t}}\left[h(\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^2)\right] = \mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right].\end{aligned}$$

Thus, we have that

$$\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^2\right] \leq \mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right]^{1/(1+\alpha/2)}$$

By taking the square root of both sides,

$$\mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^2\right]^{1/2} \leq \mathbb{E}_{\pi^*_{2:t}}\left[\left|u_k(\mathcal{H}_t^{(i)}) - l_k(\mathcal{H}_t^{(i)})\right|^{2+\alpha}\right]^{1/(2+\alpha)} \leq \eta.$$

Bracketing Number: By the above results we have that

$$N_{[\,]}\left(2\sqrt{c_2}\eta, \mathcal{G}_\partial, L_2(\mathcal{P}_{\pi^*})\right) \leq N_{[\,]}\left(\eta, \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*})\right)^2.$$

Moreover,

$$N_{[\,]}(\eta, \mathcal{G}, L_2(\mathcal{P}_{\pi^*})) \leq N_{[\,]}(\eta/(2\sqrt{c_2}), \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*}))^2.$$

Thus,

$$\int_0^1 \sqrt{\log N_{[\,]}(\eta, \mathcal{G}, L_2(\mathcal{P}_{\pi^*}))}d\eta \leq \int_0^1 \sqrt{\log N_{[\,]}(\eta/(2\sqrt{c_2}), \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*}))^2}d\eta$$

By exponent property of $\log$,

$$= \sqrt{2}\int_0^1 \sqrt{\log N_{[\,]}(\eta/(2\sqrt{c_2}), \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*}))}d\eta$$

We now use integration by substitution, with $u = \eta/(2\sqrt{c_2})$; note that $\frac{\partial u}{\partial \eta} = (2\sqrt{c_2})^{-1}$.

$$= \sqrt{2}(2\sqrt{c_2})\int_0^1 \sqrt{\log N_{[\,]}(\eta/(2\sqrt{c_2}), \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*}))}(2\sqrt{c_2})^{-1}d\eta$$

$$= \sqrt{2}(2\sqrt{c_2})\int_0^{(2\sqrt{c_2})^{-1}} \sqrt{\log N_{[\,]}(u, \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*}))}du < \infty.$$

The above is bounded by our assumption that $\int_0^1 \sqrt{\log N_{[\,]}\big(u, \mathcal{F}, L_{2+\alpha}\big(\mathcal{P}_{\pi^*}\big)\big)}du$. Note that if $(2\sqrt{c_2})^{-1} \leq 1$, the result above holds directly by this assumption. If $(2\sqrt{c_2})^{-1} > 1$, the result above holds because $N_{[\,]}\big(\eta, \mathcal{F}, L_{2+\alpha}\big(\mathcal{P}_{\pi^*}\big)\big) \leq N_{[\,]}\big(1, \mathcal{F}, L_{2+\alpha}\big(\mathcal{P}_{\pi^*}\big)\big)$ for all $\eta > 1$.

Envelope Function for $\mathcal{G}_\partial$: We can construct the envelope function for $\mathcal{G}_\partial$ using the brackets for $\mathcal{G}_\partial$ that we constructed above. Specifically, the envelope function G can be taken to be the supremum of the upper and lower bracketing functions for $\mathcal{G}_\partial$. This envelope function will be such that $\mathbb{E}_{\pi^*_{2:t}}\left[G(\mathcal{H}_t^{(i)})^2\right] < \infty$ since the brackets we constructed for $\mathcal{G}_\partial$ have finite $L_{2+\alpha}\big(\mathcal{P}_{\pi^*}\big)$ norm (see Section C.1.4 for more on the definition of bracketing functions we use). ■

C.6.1 Stochastic Equicontinuity (Lemma C.6.1)

Lemma C.6.1 (Stochastic Equicontinuity)**.** *Let $t \in [1\colon T-1]$ and let $\mathcal{F}$ be a class of real-valued, measurable functions of $\mathcal{H}_t^{(i)}$ such that $\int_0^1 \sqrt{\log N_{[\,]}\left(\varepsilon, \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*})\right)}d\varepsilon < \infty$ for some constant $\alpha > 0$. Let $\hat{f}^{(n)} \in \mathcal{F}$ be a sequence of functions such that $\nu\big(\hat{f}^{(n)}, f_0\big) \xrightarrow{p} 0$ for some $f_0 \in \mathcal{F}$ where $\nu(f,g) \triangleq \mathbb{E}_{\pi^*_{2:t}}\left[\left|f\big(\mathcal{H}_t^{(i)}\big) - g\big(\mathcal{H}_t^{(i)}\big)\right|^2\right]^{1/2}$.*

Under Conditions 5.3.2 and 5.3.3, and the condition that $\hat{\beta}^{(n)}_{1:t-1} - \beta^*_{1:t-1} = O_P(1/\sqrt{n})$ we have that

$$\mathcal{G}_{\mathcal{F}}^{(n)}\big(\hat{f}^{(n)}\big) - \mathcal{G}_{\mathcal{F}}^{(n)}\big(f_0\big) \xrightarrow{p} 0 \tag{C.6.5}$$

where

$$\mathcal{G}_{\mathcal{F}}^{(n)}(f) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right] \right).$$

Proof of Lemma C.6.1. We use an argument akin to that of Lemma 19.24 of[100],

which is for i.i.d. data. We use $\ell^\infty(\mathcal{F})$ to refer to the collection of all bounded functions from $\mathcal{F}$ to $\mathbb{R}$.

By Theorem C.6.1 (Functional Asymptotic Normality under Finite Bracketing Integral), the following stochastic process converges weakly to a mean-zero Gaussian Process $\mathcal{G}_\mathcal{F} \in \ell^\infty(\mathcal{F})$:

$$\mathcal{G}_\mathcal{F}^{(n)} \triangleq \left\{ \mathcal{G}_\mathcal{F}^{(n)}(f) \text{ s.t. } f \in \mathcal{F} \right\} \xrightarrow{D} \mathcal{G}_\mathcal{F}, \tag{C.6.6}$$

where the limit $\mathcal{G}_\mathcal{F}$ has the following covariance function:

$$\mathbb{E}\left[\mathcal{G}_\mathcal{F}(f)\mathcal{G}_\mathcal{F}(g)\right] \triangleq \mathbb{E}_{\pi^*_{2:t}}\Bigg[\left(\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]\right) \left(\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} g(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} g(\mathcal{H}_t^{(i)})\right]\right) \Bigg]. \tag{C.6.7}$$

We are able to apply Theorem C.6.1 because of our assumptions that Conditions 5.3.2 and 5.3.3 hold, and since we've assumed that $\int_0^1 \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L_{2+\alpha}(\mathcal{P}_{\pi^*}))}d\varepsilon < \infty$ and $\hat{\beta}^{(n)}_{1:t-1} - \beta^*_{1:t-1} = O_P(1/\sqrt{n})$.

By Lemma 18.15 of[99], the weak convergence result from display (C.6.6) implies that the limit $\mathcal{G}_\mathcal{F}$ can be constructed to have almost all sample paths in $\mathrm{UC}(\mathcal{F}, \rho)$, the collection of all uniformly continuous functions from $\mathcal{F}$ to $\mathbb{R}$; ρ is the standard deviation semi-metric:

$$\rho(f,g) \triangleq \sqrt{\mathbb{E}_{\pi^*_{2:t}}\left[\left|\mathcal{G}_\mathcal{F}(f) - \mathcal{G}_\mathcal{F}(g)\right|^2\right]}$$

For now, we take as given that $\rho\left(\hat{f}^{(n)}, f_0\right) \xrightarrow{P} 0$ (we show this at the end of this proof). Thus, we have that $\hat{f}^{(n)} \xrightarrow{P} f_0$. By Slutsky's theorem and the convergence result from display (C.6.6), we have that $\left(\mathcal{G}_\mathcal{F}^{(n)}, \hat{f}^{(n)}\right) \xrightarrow{D} \left(\mathcal{G}_\mathcal{F}, f_0\right)$.

Consider the evaluation function $g : \ell^\infty(\mathcal{F}) \times \mathcal{F} \mapsto \mathbb{R}$ where $g(\mathcal{G}, f) \triangleq \mathcal{G}(f) - \mathcal{G}(f_0)$. Note that the evaluation mapping g is continuous at $\left(z, \hat{f}\right) \in \ell^\infty(\mathcal{F}) \times \mathcal{F}$ if z is continuous at f; this is discussed in the proof of Lemma 18.15 of[99]. Since, as discussed earlier, $\mathcal{G}_\mathcal{F}$ can be constructed to have almost all sample paths in $\mathrm{UC}(\mathcal{F}, \rho)$, thus $\mathcal{G}_\mathcal{F}$ is at f_0 for almost all sample paths and the evaluation mapping g is continuous at $\left(\mathcal{G}_\mathcal{F}, f_0\right)$ for almost all sample paths.

Thus, by the continuous mapping theorem, we have that

$$\mathcal{G}_\mathcal{F}^{(n)}\left(\hat{f}^{(n)}\right) - \mathcal{G}_0\left(f_0\right) = g\left(\mathcal{G}_\mathcal{F}^{(n)}, \hat{f}^{(n)}\right) \xrightarrow{D} g\left(\mathcal{G}_\mathcal{F}, f_0\right) = \mathcal{G}_\mathcal{F}\left(f_0\right) - \mathcal{G}_\mathcal{F}\left(f_0\right) = 0. \tag{C.6.8}$$

Since convergence in distribution to a constant implies convergence in probability, display (C.6.8) above implies the main result display (C.6.5) holds.

We now show that $\rho\left(\hat{f}^{(n)}, f_0\right) \xrightarrow{P} 0$**.** Note that for any functions $f, g \in \mathcal{F}$,

$$\rho(f, g)^2 = \mathbb{E}_{\pi_{2:t}^*}\left[\left|\mathcal{G}_\mathcal{F}(f) - \mathcal{G}_\mathcal{F}(g)\right|^2\right]$$

$$= \mathbb{E}_{\pi_{2:t}^*}\left[\mathcal{G}_\mathcal{F}(f)^2 - 2\mathcal{G}_\mathcal{F}(f)\mathcal{G}_\mathcal{F}(g) + \mathcal{G}_\mathcal{F}(g)^2\right]$$

Using the covariance expression from display (C.6.7), we have that for $G(\mathcal{H}_t^{(i)}; f) \triangleq \left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi_{2:t}^*}\left[\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]$,

$$= \mathbb{E}_{\pi_{2:t}^*}\left[G(\mathcal{H}_t^{(i)}; f)^2 - 2G(\mathcal{H}_t^{(i)}; f)G(\mathcal{H}_t^{(i)}; g) + G(\mathcal{H}_t^{(i)}; g)^2\right]$$

$$= \mathbb{E}_{\pi_{2:t}^*}\left[\left\{G(\mathcal{H}_t^{(i)}; f) - G(\mathcal{H}_t^{(i)}; g)\right\}^2\right]$$

By the definition of $G(\mathcal{H}_t^{(i)}; f)$,

$$= \mathbb{E}_{\pi^*_{2:t}}\Bigg[\Bigg\{ \left(\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]\right) \\ - \left(\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} g(\mathcal{H}_t^{(i)}) - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} g(\mathcal{H}_t^{(i)})\right]\right) \Bigg\}^2\Bigg]$$

Let $X = \left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \left\{\pi^{*,(i)}_{2:t}\right\}^{-1} g(\mathcal{H}_t^{(i)})$. Since $\mathbb{E}_{\pi^*_{2:t}}\left[(X - \mathbb{E}_{\pi^*_{2:t}}[X])^2\right] = \mathbb{E}_{\pi^*_{2:t}}\left[X^2\right] - \mathbb{E}_{\pi^*_{2:t}}\left[X\right]^2$,

$$= \mathbb{E}_{\pi^*_{2:t}}\left[\left(\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \left\{\pi^{*,(i)}_{2:t}\right\}^{-1} g(\mathcal{H}_t^{(i)})\right)^2\right] \\ - \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} g(\mathcal{H}_t^{(i)}) - \left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]^2$$

$$\leq \mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-2} \left\{f(\mathcal{H}_t^{(i)}) - g(\mathcal{H}_t^{(i)})\right\}^2\right]$$

By Condition 5.3.2, $\left\{\pi^{*,(i)}_{2:t}\right\}^{-2} \leq \pi_{\min}^{2(t-1)}$ a.s., so

$$\leq \pi_{\min}^{2(t-1)} \mathbb{E}_{\pi^*_{2:t}}\left[\left\{f(\mathcal{H}_t^{(i)}) - g(\mathcal{H}_t^{(i)})\right\}^2\right]$$

The above implies that $\rho\left(\hat{f}^{(n)}, f_0\right) \xrightarrow{P} 0$ because by assumption of the Lemma, we have that $\nu\left(\hat{f}^{(n)}, f_0\right) \xrightarrow{P} 0$ where $\nu(f,g) \triangleq \mathbb{E}_{\pi^*_{2:t}}\left[\left|f(\mathcal{H}_t^{(i)}) - g(\mathcal{H}_t^{(i)})\right|^2\right]^{1/2}$. ■

C.7 Maximal Inequalities for Adaptively Sampled Data

Overview and Notation for Appendix C.7 Results. In this Appendix, we will use the following notation:

$$\pi^{*,(i)}_{2:t} \triangleq \prod_{t'=2}^{t} \pi^{*}_{t'}(A^{(i)}_{t'}, S^{(i)}_{t'}) \text{ and } \hat{\pi}^{(i)}_{2:t} \triangleq \prod_{t'=2}^{t} \hat{\pi}^{(n)}_{t'}(A^{(i)}_{t'}, S^{(i)}_{t'}).$$

Consider a function class $\mathcal{F}$ whose complexity is sufficiently controlled. In this section, our goal is to show a maximal inequality to bound the following:

$$\mathbb{E}^{*}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1} f(\mathcal{H}^{(i)}_{t}) - \mathbb{E}\left[\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1} f(\mathcal{H}^{(i)}_{t})\right]\right)\right|\right] \tag{C.7.1}$$

Above $\mathbb{E}^{*}$ refers to outer expectations as defined in Section 18.2[99]. A bound for the above term in display (C.7.1) is used in Theorem C.6.1 (Functional Asymptotic Normality under Finite Bracketing Integral for Adaptively Sampled Data). This maximal inequality will be a function of the bracketing integral $\mathcal{F}$ bracketing integral, $\int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon, \mathcal{F}, L_p(\mathcal{P}_{\pi^*})\big)}d\varepsilon$, which we assume is finite.

Note that since $\{\mathcal{H}^{(i)}_{T}\}_{i=1}^{n}$ are not independent in our setting, we cannot use classical maximal inequalities for i.i.d. data to bound the term in display (C.7.1). See Section 19.6 of[99] for information on maximal inequalities for i.i.d. data. Our results build on the ideas used in these results for i.i.d. data.

Summary of Results in this Section

- **Weighted Martingale Bernstein Inequality (Lemma C.7.2)** proves a Bernstein inequality for our non-independent, adaptively sampled data type and is the most

novel step in this section. The proof leverages the conditional independence of the action selection at each time-step and the fact that the underlying potential outcomes are i.i.d. The proof repeatedly uses a key helper Lemma C.7.1 (Moving Products out of Expectations using Weights).

- **Maximal Inequality for Finite Class of Functions (Lemma C.7.3)** proves a maximal inequality to bound the term in display (C.7.1) in the case that $|\mathcal{F}| < \infty$. The proof closely follows that of Lemma 19.33[99], but replaces the use of a Bernstein inequality for i.i.d. data with Lemma C.7.2 (Weighted Martingale Bernstein Inequality).

- **Maximal Inequality as a Function of the Bracketing Integral (Lemma C.8.1)** proves a maximal inequality to bound the term in display (C.7.1) as a function of the bracketing integral for $\mathcal{F}$. The proof closely follows that of Lemma 19.34[99], but replaces the use of a maximal inequality for empirical processes for a finite class of functions on i.i.d. data with Lemma C.7.3 (Maximal Inequality for Finite Class of Functions).

C.7.1 Moving Products out of Expectations using Weights (Lemma C.7.1)

Lemma C.7.1 (Moving Products out of Expectations using Weights)**.** *For any $t \in [2 : T]$, let f be any real-valued, measurable function of $\mathcal{H}_t^{(i)}$ such that $\mathbb{E}_{\pi^*_{2:t}}\left[\left|f(\mathcal{H}_t^{(i)})\right|\right] < \infty$. Let c be fixed constants. The following equality holds for any $n \geq 1$:*

$$\mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) + c\right)\right] = \prod_{i=1}^{n}\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) + c\right]. \tag{C.7.2}$$

Remark C.7.1 (Display (C.7.2) Comment)**.** *Note that regarding the expectation terms on the right hand side above, for any stochastic policies* $\pi_{2:t}(\beta_{1:t-1})$*,*

$$\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}^{(i)}_t)\right] = \mathbb{E}_{\pi^*_{2:t}}\left[\left(\prod_{t'=2}^{t} \pi^*_{t'}(A^{(i)}_{t'}, S^{(i)}_{t'})\right)^{-1} f(\mathcal{H}^{(i)}_t)\right]$$
$$= \mathbb{E}_{\pi_{2:t}(\beta_{1:t-1})}\left[\left(\prod_{t'=2}^{t} \pi_{t'}(A^{(i)}_{t'}, S^{(i)}_{t'}; \beta_{t'-1})\right)^{-1} f(\mathcal{H}^{(i)}_t)\right].$$

Proof of Lemma C.7.1 (Conditional Independence using Weights). For notational convenience we consider the t set to T case; the argument holds by the same argument for any $t \in [2\colon T]$.

Let $t \in [2 : T]$ and let g be a real-valued, measurable function of $\mathcal{H}^{(i)}_t, S^{(i)}_{t+1}$. A key result which we will take as given for now (we prove it at the end of this proof), is that

$$\mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1} g\left(\mathcal{H}^{(i)}_t, S^{(i)}_{t+1}\right) + c\right)\right]$$
$$= \mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}^{(i)}_{2:t-1}\right\}^{-1} \mathbb{E}_{\pi^*_t}\left[\left\{\pi^{*,(i)}_t\right\}^{-1} g\left(\mathcal{H}^{(i)}_t, S^{(i)}_{t+1}\right)\middle| \mathcal{H}^{(i)}_{t-1}, S^{(i)}_t\right] + c\right)\right]. \quad \text{(C.7.3)}$$

We now show that the desired result holds by repeatedly applying display (C.7.3). Applying display (C.7.3) for t set to T and for g set to f, we have that

$$\mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}^{(i)}_T) + c\right)\right]$$

$$= \mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}^{(i)}_{2:T-1}\right\}^{-1} \mathbb{E}_{\pi^*_T}\left[\left\{\pi^{*,(i)}_T\right\}^{-1} f(\mathcal{H}^{(i)}_t)\middle| \mathcal{H}^{(i)}_{T-1}, S^{(i)}_T\right] + c\right)\right]$$

Now, note that $\mathbb{E}_{\pi_T^*}\left[\left\{\pi_T^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) \middle| \mathcal{H}_{T-1}^{(i)}, S_T^{(i)}\right]$ is a function of $\mathcal{H}_{T-1}^{(i)}, S_T^{(i)}$; let this be function be g when we apply display (C.7.3) again for t set to $T-1$.

$$= \mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}_{2:T-2}^{(i)}\right\}^{-1}\mathbb{E}_{\pi_{T-1}^*}\left[\left\{\pi_{T-1}^{*,(i)}\right\}^{-1}\mathbb{E}_{\pi_T^*}\left[\left\{\pi_T^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) \middle| \mathcal{H}_{T-1}^{(i)}, S_T^{(i)}\right] \middle| \mathcal{H}_{T-2}^{(i)}, S_{T-1}^{(i)}\right] + c\right)\right]$$

Since $\left\{\pi_{T-1}^{*,(i)}\right\}^{-1}$ is a constant given $\mathcal{H}_{T-1}^{(i)}$,

$$= \mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}_{2:T-2}^{(i)}\right\}^{-1}\mathbb{E}_{\pi_{T-1}^*}\left[\mathbb{E}_{\pi_T^*}\left[\left\{\pi_{T-1:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) \middle| \mathcal{H}_{T-1}^{(i)}, S_T^{(i)}\right] \middle| \mathcal{H}_{T-2}^{(i)}, S_{T-1}^{(i)}\right] + c\right)\right]$$

By law of iterated expectations,

$$= \mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}_{2:T-2}^{(i)}\right\}^{-1}\mathbb{E}_{\pi_{T-1:T}^*}\left[\left\{\pi_{T-1:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) \middle| \mathcal{H}_{T-2}^{(i)}, S_{T-1}^{(i)}\right] + c\right)\right].$$

By repeatedly applying display (C.7.3) and the above argument for t set to $T-2, T-3, \ldots, 2$ we have that

$$= \mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}_{2:T-3}^{(i)}\right\}^{-1}\mathbb{E}_{\pi_{T-2:T}^*}\left[\left\{\pi_{T-2:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) \middle| \mathcal{H}_3^{(i)}, S_{T-2}^{(i)}\right] + c\right)\right]$$

$$= \mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}_{2:T-4}^{(i)}\right\}^{-1}\mathbb{E}_{\pi_{T-3:T}^*}\left[\left\{\pi_{T-3:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) \middle| \mathcal{H}_4^{(i)}, S_{T-3}^{(i)}\right] + c\right)\right]$$

$$= \ldots = \mathbb{E}\left[\prod_{i=1}^{n}\left(\mathbb{E}_{\pi_{2:T}^*}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) \middle| \mathcal{H}_1^{(i)}, S_1^{(i)}\right] + c\right)\right].$$

Finally, recall that $\left\{\mathcal{H}_1^{(i)}, S_2^{(i)}\right\}_{i=1}^n = \left\{S_1^{(i)}, A_1^{(i)}, Y_1^{(i)}, S_2^{(i)}\right\}_{i=1}^n$ are independent over $i \in$

$[1:n]$. Thus,

$$\mathbb{E}\left[\prod_{i=1}^{n}\left\{\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}^{(i)}_T)\middle|\mathcal{H}^{(i)}_1, S^{(i)}_1\right] + c\right\}\right]$$
$$= \prod_{i=1}^{n}\mathbb{E}\left[\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}^{(i)}_T)\middle|\mathcal{H}^{(i)}_1, S^{(i)}_1\right] + c\right]$$

By law of iterated expectations,

$$= \prod_{i=1}^{n}\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}^{(i)}_T) + c\right].$$

Thus we have shown that the desired result holds and all that is left is to show that display (C.7.3) holds.

Proof of display (C.7.3)**.** The proof of display (C.7.3) leverages (i) the Radon-Nikodym weights and (ii) conditional independence properties. Pick any $t \in [2:T]$ and let g be a real-valued, measurable function of $\mathcal{H}^{(i)}_t, S^{(i)}_{t+1}$. By law of iterated expectations,

$$\mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1} g\left(\mathcal{H}^{(i)}_t, S^{(i)}_{t+1}\right) + c\right)\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1} g\left(\mathcal{H}^{(i)}_t, S^{(i)}_{t+1}\right) + c\right)\middle|\mathcal{H}^{(1:n)}_{t-1}, S^{(1:n)}_t\right]\right]$$

Note that the conditional expectation $\mathbb{E}\left[\prod_{i=1}^{n}\left(\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1} g\left(\mathcal{H}^{(i)}_t, S^{(i)}_{t+1}\right) + c\right)\middle|\mathcal{H}^{(1:n)}_{t-1}, S^{(1:n)}_t\right]$ is only integrating over $\left\{A^{(i)}_t, Y^{(i)}_t, S^{(i)}_{t+1}\right\}_{i=1}^{n}$ (the policy parameters $\hat{\beta}^{(n)}_{1:t-1}$ used in $\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1}$ are known given $\mathcal{H}^{(1:n)}_{t-1}$). Additionally, note that conditional on $\mathcal{H}^{(1:n)}_{t-1}, S^{(1:n)}_t$,

$\{A_t^{(i)}, Y_t^{(i)}, S_{t+1}^{(i)}\}$ are independent over $i \in [1\colon n]$. Thus,

$$= \mathbb{E}\Bigg[\prod_{i=1}^{n} \Bigg(\mathbb{E}\left[\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} g\big(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\big) \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] + c \Bigg) \Bigg]$$

Since $\left\{ \hat{\pi}_{2:t-1}^{(i)} \right\}^{-1} = \left(\prod_{t'=2}^{t-1} \hat{\pi}_{t'}^{(n)}(A_{t'}^{(i)}, S_{t'}^{(i)}) \right)^{-1}$ is a constant given $\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}$,

$$= \mathbb{E}\Bigg[\prod_{i=1}^{n} \Bigg(\left\{ \hat{\pi}_{2:t-1}^{(i)} \right\}^{-1} \mathbb{E}\left[\left\{ \hat{\pi}_t^{(i)} \right\}^{-1} g\big(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\big) \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] + c \Bigg) \Bigg]$$

Now now show that the above equals the following:

$$= \mathbb{E}\Bigg[\prod_{i=1}^{n} \Bigg(\left\{ \hat{\pi}_{2:t-1}^{(i)} \right\}^{-1} \mathbb{E}_{\pi_t^*}\left[\left\{ \pi_t^{*,(i)} \right\}^{-1} g\big(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\big) \middle| \mathcal{H}_{T-1}^{(i)}, S_t^{(i)} \right] + c \Bigg) \Bigg]. \tag{C.7.4}$$

Showing the above will be sufficient for display (C.7.3).

Note that since $1 = \pi_t^*(A_t^{(i)}, S_t^{(i)})^{-1} \pi_t^*(A_t^{(i)}, S_t^{(i)}) = \left\{ \pi_t^{*,(i)} \right\}^{-1} \pi_t^*(A_t^{(i)}, S_t^{(i)})$,

$$\begin{aligned}
&\mathbb{E}\left[\left\{ \hat{\pi}_t^{(i)} \right\}^{-1} g\big(\mathcal{H}_T^{(i)}, S_{t+1}^{(i)}\big) \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] \\
&\qquad = \mathbb{E}\left[\left\{ \hat{\pi}_t^{(i)} \right\}^{-1} \left\{ \pi_t^{*,(i)} \right\}^{-1} \pi_t^*(A_t^{(i)}, S_t^{(i)}) g\big(\mathcal{H}_T^{(i)}, S_{t+1}^{(i)}\big) \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right] \\
&\qquad\qquad = \mathbb{E}_{\pi_t^*}\left[\left\{ \pi_t^{*,(i)} \right\}^{-1} g\big(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\big) \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right].
\end{aligned}$$

The last equality above holds because $\left\{ \hat{\pi}_t^{(i)} \right\}^{-1} \pi_t^*(A_t^{(i)}, S_t^{(i)}) = W_t^{(i)}(\beta_{t-1}^*, \hat{\beta}_{t-1}^{(n)})$.

Also, note that the expectation $\mathbb{E}_{\pi_t^*}\left[\left\{ \pi_t^{*,(i)} \right\}^{-1} g\big(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\big) \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)} \right]$ integrates over $\{A_t^{(i)}, Y_t^{(i)}, S_{t+1}^{(i)}\}$. Since actions are selected using π_t^* rather than $\hat{\pi}_t^{(n)}$ in the expectation, the distribution of $\{A_t^{(i)}, Y_t^{(i)}, S_{t+1}^{(i)}\}$ depends only on $\mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}$ through $\mathcal{H}_{t-1}^{(i)}, S_t^{(i)}$.

This means that

$$\mathbb{E}_{\pi_t^*}\left[\left\{\pi_t^{*,(i)}\right\}^{-1} g\left(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\right) \middle| \mathcal{H}_{t-1}^{(1:n)}, S_t^{(1:n)}\right] = \mathbb{E}_{\pi_t^*}\left[\left\{\pi_t^{*,(i)}\right\}^{-1} g\left(\mathcal{H}_t^{(i)}, S_{t+1}^{(i)}\right) \middle| \mathcal{H}_{t-1}^{(i)}, S_t^{(i)}\right].$$

Thus we have shown display (C.7.4) holds. ■

C.7.2 Weighted Martingale Bernstein Inequality (Lemma C.7.2)

Lemma C.7.2 (Weighted Martingale Bernstein Inequality)**.** *We assume Condition 5.3.2 (Minimum Exploration) holds. Let f be a real-valued, measurable function of $\mathcal{H}_t^{(i)}$. Then, for any $x > 0$ and for all $n \geq 1$,*

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]\right| \geq x\right)$$
$$\leq 2\exp\left(-\frac{\pi_{\min}^{t-1}}{4}\frac{x^2}{\mathbb{E}_{\pi_{2:t}^*}\left[\left\{\pi_{2:t}^{*,(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})^2\right] + x\|f\|_\infty/\sqrt{n}}\right). \quad \text{(C.7.5)}$$

Above $\|f\|_\infty \triangleq \sup_h |f(h)|$.

Proof of Lemma C.7.2 (Weighted Martingale Bernstein Inequality). We follow an argument similar to Lemma 19.32 in [99]. For notational convenience we consider the t set to T case; the argument holds by the same argument for any $t \in [2\colon T]$.

The leading 2 in display (C.7.5) is due to separate bounds for the upper and lower tail bounds. It is sufficient to show the upper tail bound, because the lower tail bound holds by the upper tail bound applied to $-f$.

Note if $\|f\|_\infty = 0$, then $\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} f(\mathcal{H}_t^{(i)})\right]\right) \geq x\right)$

$= 0$, so in this case display (C.7.5) easily holds. Thus, for the remainder of the proof we assume that $\|f\|_\infty > 0$.

Let $x > 0$.

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) \right] \right) \geq x \right) \tag{C.7.6}$$

Using a Chernoff bound, for any $\lambda > 0$,

$$\leq e^{-\lambda x} \mathbb{E}\left[\exp\left\{ \frac{\lambda}{\sqrt{n}} \sum_{i=1}^{n} \left(\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) \right] \right) \right\} \right]$$

Using properties of exponents, we can change the summation in exponent into a product,

$$= e^{-\lambda x} \mathbb{E}\left[\prod_{i=1}^{n} \exp\left\{ \frac{\lambda}{\sqrt{n}} \left(\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) \right] \right) \right\} \right]$$

We now apply Maclaurin series for exponential function, i.e., that $e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}$.

$$= e^{-\lambda x} \mathbb{E}\left[\prod_{i=1}^{n} \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k \left(\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) \right] \right)^k \right]$$

By simplifying the first two terms in the inner summation,

$$\begin{aligned} = e^{-\lambda x} \mathbb{E}\Bigg[\prod_{i=1}^{n} \Bigg\{ 1 + \frac{\lambda}{\sqrt{n}} \left(\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) \right] \right) \\ + \sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k \left(\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{ \hat{\pi}_{2:T}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) \right] \right)^k \Bigg\} \Bigg] \end{aligned} \tag{C.7.7}$$

Now note the following observations:

- Let $\varepsilon > 0$. By Condition 5.3.2 (Minimum Exploration),

$$\left\{\hat{\pi}_{2:t}^{(i)}\right\}^{-1} = \left[\prod_{t'=2}^{t} \hat{\pi}_{t'}^{(n)}(A_{t'}^{(i)}, S_{t'}^{(i)})\right]^{-1} \leq \pi_{\min}^{-(t-1)} \text{ a.s.,} \tag{C.7.8}$$

and

$$\begin{aligned}&\left|\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right| \\ &\qquad \leq \left|\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right| + \left|\mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right| \leq 2\pi_{\min}^{-(T-1)} \|f\|_\infty. \end{aligned} \tag{C.7.9}$$

- We can upper bound the following:

$$\left(\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^2$$

$$= (\hat{\pi}_{2:T}^{(i)})^{-2} f(\mathcal{H}_T^{(i)})^2 - 2\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)}) \mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right] + \left(\mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^2$$

Note that $\mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right] = \mathbb{E}_{\pi_{2:T}^*}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]$.

$$\begin{aligned}= \left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} &\left(\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})^2 - 2f(\mathcal{H}_T^{(i)}) \mathbb{E}_{\pi_{2:T}^*}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right) \\ &+ \left(\mathbb{E}_{\pi_{2:T}^*}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^2\end{aligned}$$

Since $\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} \leq \pi_{\min}^{-(T-1)}$ a.s. by display (C.7.8),

$$\leq \left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1}\Bigg(\underbrace{\pi_{\min}^{-(T-1)} f(\mathcal{H}_T^{(i)})^2 - 2f(\mathcal{H}_T^{(i)})\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]}_{\triangleq g(\mathcal{H}_T^{(i)})}\Bigg) + \left(\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^2$$

For $g(\mathcal{H}_T^{(i)}) \triangleq \pi_{\min}^{-(T-1)} f(\mathcal{H}_T^{(i)})^2 - 2f(\mathcal{H}_T^{(i)})\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]$,

$$= \left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} g(\mathcal{H}_T^{(i)}) + \left(\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^2.$$

Thus, in summary we have that

$$\left(\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^2 \leq \left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} g(\mathcal{H}_T^{(i)}) + \left(\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^2. \quad \text{(C.7.10)}$$

- By display (C.7.9) and (C.7.10), we have that for any $k \geq 2$

$$\left(\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^k$$

$$\leq \left(\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right)^2 \left|\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)}) - \mathbb{E}\left[\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})\right]\right|^{k-2}$$

$$\leq \left(\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} g(\mathcal{H}^{(i)}_T) + \left(\mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right] \right)^2 \right) \left\{ 2\pi_{\min}^{-(T-1)} \|f\|_\infty \right\}^{k-2} \tag{C.7.11}$$

Note that in display (C.7.7), each of the terms in the product over n terms is non-negative because the i^{th} term in the product equals $\exp\left\{ \frac{\lambda}{\sqrt{n}} \left(\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) - \mathbb{E}\left[\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right] \right) \right\}$ and $e^x \geq 0$ for all x. Thus, by display (C.7.11), we can upper bound display (C.7.7) as follows:

$$\begin{aligned} &\leq e^{-\lambda x} \mathbb{E}\Bigg[\prod_{i=1}^{n} \Bigg\{ 1 + \frac{\lambda}{\sqrt{n}} \left(\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) - \mathbb{E}\left[\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right] \right) \\ &+ \sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k \left(\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} g(\mathcal{H}^{(i)}_T) + \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2 \right) \left(2\pi_{\min}^{-(T-1)} \|f\|_\infty \right)^{k-2} \Bigg\} \Bigg] \end{aligned} \tag{C.7.12}$$

Note that everything in the expectation above in display (C.7.12) is bounded a.s.; we will show that this is true for the infinite summation over k. Let $y = \left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} g(\mathcal{H}^{(i)}_T) + \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2$ and $z = 2\pi_{\min}^{-(T-1)} \|f\|_\infty$. Note that both y and z are bounded a.s. Thus, since $\|f\|_\infty > 0$ by assumption (we discussed the $\|f\|_\infty = 0$ case at the beginning of this proof), we have that $z > 0$, so $\sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k yz^{k-2} = yz^{-2} \sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k z^k = yz^{-2} \left(e^{z\lambda/\sqrt{n}} - \sum_{k=0}^{1} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k z^k \right)$ is also bounded a.s.

Moreover, display (C.7.12) can be written as $e^{-\lambda x} \mathbb{E}\left[\prod_{i=1}^{n} \left(\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} h(\mathcal{H}^{(i)}_T) + c \right) \right]$, for some function h and some finite constant c. (Note that $\mathbb{E}\left[\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right] = \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]$) Thus, we can apply Lemma C.7.1 (Moving Products out of

Expectations using Weights) to get that display (C.7.12) is equal to

$$= e^{-\lambda x} \prod_{i=1}^{n} \mathbb{E}_{\pi^*_{2:T}} \Bigg[1 + \frac{\lambda}{\sqrt{n}} \Bigg(\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) - \mathbb{E} \left[\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right] \Bigg)$$
$$+ \sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k \left(\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} g(\mathcal{H}^{(i)}_T) + \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2 \right) \left(2\pi_{\min}^{-(T-1)} \|f\|_\infty \right)^{k-2} \Bigg].$$

Since $\mathbb{E} \left[\left\{ \hat{\pi}^{(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right] = \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]$, we can cancel terms in the first line above.

$$= e^{-\lambda x} \prod_{i=1}^{n} \Bigg\{ 1 + \mathbb{E}_{\pi^*_{2:T}} \Bigg[\sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k \left(\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} g(\mathcal{H}^{(i)}_T) + \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2 \right)$$
$$\left(2\pi_{\min}^{-(T-1)} \|f\|_\infty \right)^{k-2} \Bigg] \Bigg\}$$

Since everything in the expectations above are bounded a.s. (discussed below display (C.7.12)), we can exchange the expectation with the infinite summation over k.

$$= e^{-\lambda x} \prod_{i=1}^{n} \Bigg\{ 1 + \sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k \left(\mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} g(\mathcal{H}^{(i)}_T) \right] + \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2 \right)$$
$$\left(2\pi_{\min}^{-(T-1)} \|f\|_\infty \right)^{k-2} \Bigg\}. \quad \text{(C.7.13)}$$

Note the following:

- Recall that $g(\mathcal{H}^{(i)}_T) \triangleq \pi_{\min}^{-(T-1)} f(\mathcal{H}^{(i)}_T)^2 - 2f(\mathcal{H}^{(i)}_T) \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]$. So,

$$\mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} g(\mathcal{H}^{(i)}_T) \right] + \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2$$

$$= \pi_{\min}^{-(T-1)} \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T)^2 \right] - 2\mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2 + \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2$$

$$= \pi_{\min}^{-(T-1)} \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T)^2 \right] - \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T) \right]^2$$

$$\leq \pi_{\min}^{-(T-1)} \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T)^2 \right]$$

Thus display (C.7.13) can be upper bounded by the following:

$$e^{-\lambda x} \prod_{i=1}^{n} \left\{ 1 + \sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k \pi_{\min}^{-(T-1)} \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T)^2 \right] \left(2\pi_{\min}^{-(T-1)} \|f\|_\infty \right)^{k-2} \right\}$$

By i.i.d. potential outcomes,

$$= e^{-\lambda x} \left\{ 1 + \sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}} \right)^k \pi_{\min}^{-(T-1)} \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T)^2 \right] \left(2\pi_{\min}^{-(T-1)} \|f\|_\infty \right)^{k-2} \right\}^n$$

By rearranging terms,

$$= e^{-\lambda x} \left\{ 1 + \frac{1}{n} \sum_{k=2}^{\infty} \frac{1}{k!} \frac{1}{2} \lambda^k \underbrace{2\pi_{\min}^{-(T-1)} \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}^{(i)}_T)^2 \right]}_{\triangleq \lambda_1^{-1}} \left(\underbrace{2\pi_{\min}^{-(T-1)} \|f\|_\infty / \sqrt{n}}_{\triangleq \lambda_2^{-1}} \right)^{k-2} \right\}^n$$

$$= e^{-\lambda x} \left\{ 1 + \frac{1}{n} \sum_{k=2}^{\infty} \frac{1}{k!} \frac{1}{2} \lambda^k \left(\lambda_1^{-1} \lambda_2^{-(k-2)} \right) \right\}^n \tag{C.7.14}$$

Note that since $\lambda_1^{-1}, \lambda_2^{-1} \geq 0$,

$$\lambda \triangleq x \left(\lambda_1^{-1} + x\lambda_2^{-1} \right)^{-1} \leq \min \left\{ x \left(\lambda_1^{-1} + 0 \right)^{-1}, x \left(0 + x\lambda_2^{-1} \right)^{-1} \right\} = \min \left(x\lambda_1, \lambda_2 \right).$$

Thus we have that $\lambda^k \leq \lambda \min(x\lambda_1, \lambda_2)^{k-1} \leq \lambda x \lambda_1 \lambda_2^{k-2}$. So we can upper bound display (C.7.14) as follows:

$$\leq e^{-\lambda x}\left\{1 + \frac{1}{n}\sum_{k=2}^{\infty}\frac{1}{k!}\frac{1}{2}\left(\lambda x \lambda_1 \lambda_2^{k-2}\right)\left(\lambda_1^{-1}\lambda_2^{-(k-2)}\right)\right\}^n$$

$$= e^{-\lambda x}\left\{1 + \frac{1}{n}\underbrace{\sum_{k=2}^{\infty}\frac{1}{k!}}_{\leq 1}\frac{1}{2}\lambda x\right\}^n$$

By the Maclaurin series for exponential function, $e^z = \sum_{k=0}^{\infty}\frac{z^k}{k!}$, we have $\sum_{k=2}^{\infty}\frac{1}{k!} = e - \frac{1}{0!} - \frac{1}{1!} = e - 2 \leq 1$.

$$\leq e^{-\lambda x}\left\{1 + \frac{1}{n}\frac{1}{2}x\lambda\right\}^n$$

Again by the Maclaurin series for exponential function, $e^z = \sum_{k=0}^{\infty}\frac{z^k}{k!}$, so for $z > 0$ we have that $1 + z \leq e^z$. This means that $(1+z)^n \leq e^{zn}$.

$$\leq e^{-\lambda x}\exp\left(\frac{1}{2}x\lambda\right) = \exp\left(-\frac{1}{2}x\lambda\right)$$

Recall that $\lambda \triangleq x\left(\lambda_1^{-1} + x\lambda_2^{-1}\right)^{-1}$, so,

$$= \exp\left(-\frac{1}{2}xx\left(\lambda_1^{-1} + x\lambda_2^{-1}\right)^{-1}\right)$$

Recall that $\lambda_1^{-1} = 2\pi_{\min}^{-(T-1)}\mathbb{E}_{\pi_{2:T}^*}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})^2\right]$ and $\lambda_2^{-1} = 2\pi_{\min}^{-(T-1)}\|f\|_\infty/\sqrt{n}$. So,

$$= \exp\left(-\frac{\pi_{\min}^{T-1}}{4}\frac{x^2}{\mathbb{E}_{\pi_{2:T}^*}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1} f(\mathcal{H}_T^{(i)})^2\right] + x\|f\|_\infty/\sqrt{n}}\right). \blacksquare$$

C.7.3 Maximal Inequality for Finite Class of Functions (Lemma C.7.3)

Lemma C.7.3 (Maximal Inequality for Finite Class of Functions)**.** *Let $\mathcal{F}$ be a finite class of bounded, real-valued, measurable functions of $\mathcal{H}_t^{(i)}$ with size $|\mathcal{F}| \geq 2$. For $f \in \mathcal{F}$ we define*

$$\mathcal{G}_n(f) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} f(\mathcal{H}_t^{(i)}) \right] \right).$$

Under Condition 5.3.2 (Minimum Exploration) for all sufficiently large n,

$$\mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \mathcal{G}_n(f) \right| \right] \leq C \Bigg\{ \pi_{\min}^{-(T-1)} \max_{f \in \mathcal{F}} \frac{\|f\|_\infty}{\sqrt{n}} \log(|\mathcal{F}|) + \sqrt{\pi_{\min}^{-(t-1)}} \max_{f \in \mathcal{F}} \sqrt{\mathbb{E}_{\pi_{2:t}^*}\left[\left\{ \pi_{2:t}^{*,(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)})^2 \right]} \sqrt{\log(|\mathcal{F}|)} \Bigg\}, \quad \text{(C.7.15)}$$

for some universal positive constant C (specified in the proof).

Proof of Lemma C.7.3 (Maximal Inequality for Finite Class of Functions). Our proof follows a very similar argument to Lemma 19.33 in [99]. Specifically, our proof only deviates because we use our Lemma C.7.2 (Weighted Martingale Bernstein Inequality) to prove displays (C.7.19) and (C.7.21) below.

For notational convenience we consider the t set to T case; the argument holds by the same argument for any $t \in [2 : T]$.

Special cases.

- Note that if $f \in \mathcal{F}$ such that $\|f\|_\infty = 0$, then $\mathcal{G}_n(f) = 0$. These zero functions do not contribute to increasing the upper bound for $\mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \mathcal{G}_n(f) \right| \right]$. Thus, we assume that $\|f\|_\infty > 0$ for all $f \in \mathcal{F}$ for the remainder of this proof, as this is the

most difficult case.

- Note that $f \in \mathcal{F}$ such that $\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}^{(i)}_T)^2\right] = 0$, then $f(\mathcal{H}^{(i)}_T)$ is a constant function and $\mathcal{G}_n(f) = 0$. These constant functions do not contribute to increasing the upper bound for $\mathbb{E}\left[\max_{f\in\mathcal{F}} \left|\mathcal{G}_n(f)\right|\right]$. Thus, we assume that $\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}^{(i)}_T)^2\right] > 0$ for all $f \in \mathcal{F}$ for the remainder of this proof, as this is the most difficult case.

Main argument. Let u, v be non-negative, real-valued functions of $f \in \mathcal{F}$ such that

- $u(f) = 24\pi_{\min}^{-(T-1)}\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}^{(i)}_T)^2\right]$
- $v(f) = 24\pi_{\min}^{-(T-1)}\|f\|_\infty/\sqrt{n}$

Note that

$$\mathbb{E}\left[\max_{f\in\mathcal{F}} |\mathbb{G}_n(f)|\right] = \mathbb{E}\left[\max_{f\in\mathcal{F}}\left\{|\mathbb{G}_n(f)|\mathbb{I}_{|\mathbb{G}_n(f)|>u(f)/v(f)} + |\mathbb{G}_n(f)|\mathbb{I}_{|\mathbb{G}_n(f)|\le u(f)/v(f)}\right\}\right]$$

$$\le \mathbb{E}\left[\max_{f\in\mathcal{F}} |\mathbb{G}_n(f)|\mathbb{I}_{|\mathbb{G}_n(f)|>u(f)/v(f)}\right] + \mathbb{E}\left[\max_{f\in\mathcal{F}} |\mathbb{G}_n(f)|\mathbb{I}_{|\mathbb{G}_n(f)|\le u(f)/v(f)}\right]$$

Let $\underline{\mathcal{G}_n}(f) \triangleq |\mathbb{G}_n(f)|\mathbb{I}_{|\mathbb{G}_n(f)|>u(f)/v(f)}$ and $\overline{\mathcal{G}_n}(f) \triangleq |\mathbb{G}_n(f)|\mathbb{I}_{|\mathbb{G}_n(f)|\le u(f)/v(f)}$.

$$= \mathbb{E}\left[\max_{f\in\mathcal{F}} \underline{\mathcal{G}_n}(f)\right] + \mathbb{E}\left[\max_{f\in\mathcal{F}} \overline{\mathcal{G}_n}(f)\right]$$

$$\le \mathbb{E}\left[\max_{f\in\mathcal{F}} \underline{\mathcal{G}_n}(f)/v(f)\right]\left(\max_{f\in\mathcal{F}} v(f)\right) + \mathbb{E}\left[\max_{f\in\mathcal{F}} \overline{\mathcal{G}_n}(f)/\sqrt{u(f)}\right]\left(\max_{f\in\mathcal{F}} \sqrt{u(f)}\right) \quad \text{(C.7.16)}$$

The main results show in this proof are the following:

$$\mathbb{E}\left[\max_{f\in\mathcal{F}} \underline{\mathcal{G}_n}(f)/v(f)\right] \leq \log\left(1+|\mathcal{F}|\right) \tag{C.7.17}$$

$$\mathbb{E}\left[\max_{f\in\mathcal{F}} \overline{\mathcal{G}_n}(f)/\sqrt{u(f)}\right] \leq \sqrt{\log\left(1+|\mathcal{F}|\right)} \tag{C.7.18}$$

For now we take displays (C.7.17) and (C.7.18) as given and show why the Lemma holds. Using these two results, we have that display (C.7.16) can be upper bounded by the following:

$$\begin{aligned}
&\leq \log\left(1+|\mathcal{F}|\right)\left(\max_{f\in\mathcal{F}} \underbrace{24\pi_{\min}^{-(T-1)}\|f\|_\infty/\sqrt{n}}_{v(f)}\right) \\
&\qquad + \sqrt{\log\left(1+|\mathcal{F}|\right)}\left(\max_{f\in\mathcal{F}} \underbrace{\sqrt{24\pi_{\min}^{-(T-1)}\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})^2\right]}}_{\sqrt{u(f)}}\right)
\end{aligned}$$

$$\begin{aligned}
&= 24\pi_{\min}^{-(T-1)}\left(\max_{f\in\mathcal{F}}\frac{\|f\|_\infty}{\sqrt{n}}\right)\log\left(1+|\mathcal{F}|\right) \\
&\qquad + \sqrt{24\pi_{\min}^{-(T-1)}}\left(\max_{f\in\mathcal{F}}\sqrt{\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})^2\right]}\right)\sqrt{\log\left(1+|\mathcal{F}|\right)}
\end{aligned}$$

Since $c_{\log} \triangleq \sup_{x \geq 2} \frac{\log(1+x)}{\log(x)}$ is bounded, $\frac{\log(1+|\mathcal{F}|)}{\log(|\mathcal{F}|)} \leq c_{\log}$ so, $1 \leq c_{\log} \frac{\log(|\mathcal{F}|)}{\log(1+|\mathcal{F}|)}$.

$$\leq c_{\log} 24 \pi_{\min}^{-(T-1)} \left(\max_{f \in \mathcal{F}} \frac{\|f\|_\infty}{\sqrt{n}} \right) \log\left(|\mathcal{F}|\right) + \sqrt{c_{\log} 24 \pi_{\min}^{-(T-1)}} \left(\max_{f \in \mathcal{F}} \sqrt{\mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}_T^{(i)})^2 \right]} \sqrt{\log\left(|\mathcal{F}|\right)} \right)$$

$$\leq 24 \max\left(c_{\log}, c_{\log}^{1/2}\right) \left\{ \pi_{\min}^{-(T-1)} \left(\max_{f \in \mathcal{F}} \frac{\|f\|_\infty}{\sqrt{n}} \right) \log\left(|\mathcal{F}|\right) + \sqrt{\pi_{\min}^{-(T-1)}} \left(\max_{f \in \mathcal{F}} \sqrt{\mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}_T^{(i)})^2 \right]} \right) \sqrt{\log\left(|\mathcal{F}|\right)} \right\}.$$

The above implies that the desired result, display (C.7.15), holds. All that remains is to prove that displays (C.7.17) and (C.7.18) hold.

Proving display (C.7.17) ***holds.*** Let $x > 0$. We now state some results and discuss why they hold below. For all $n \geq 1$ we have that for any $f \in \mathcal{F}$,

$$\mathbb{P}\left(\left|\underline{\mathcal{G}}_n(f)\right| \geq x\right) \underbrace{\leq}_{(a)} \mathbb{P}\left(\left|\underline{\mathcal{G}}_n(f)\right| \geq \max\left\{x, u(f)/v(f)\right\}\right)$$
$$\underbrace{\leq}_{(b)} 2\exp\left(-6 \frac{\max\left\{x, u(f)/v(f)\right\}^2}{u(f) + \max\left\{x, u(f)/v(f)\right\} v(f)} \right) \underbrace{\leq}_{(c)} 2\exp\left(-3\frac{x}{v(f)}\right). \quad \text{(C.7.19)}$$

- Inequality (a) holds because recall that $\underline{\mathcal{G}_n}(f) \triangleq |\mathbb{G}_n(f)| \mathbb{I}_{|\mathbb{G}_n(f)| > u(f)/v(f)}$.
- Inequality (b) holds by Lemma C.7.2 (Weighted Martingale Bernstein Inequality) since Condition 5.3.2 holds. Recall that $u(f) = 24\pi_{\min}^{-(T-1)} \mathbb{E}_{\pi^*_{2:T}} \left[\left\{ \pi^{*,(i)}_{2:T} \right\}^{-1} f(\mathcal{H}_T^{(i)})^2 \right]$ and $v(f) = 24\pi_{\min}^{-(T-1)} \|f\|_\infty / \sqrt{n}$.

- Inequality (c) holds because

$$6\frac{\max\left\{x, u(f)/v(f)\right\}^2}{u(f) + \max\left\{x, u(f)/v(f)\right\}v(f)} = 6\frac{\max\left\{x, u(f)/v(f)\right\}}{u(f)/\max\left\{x, u(f)/v(f)\right\} + v(f)}$$
$$\geq 6\frac{x}{u(f)/\max\left\{x, u(f)/v(f)\right\} + v(f)} \geq 6\frac{x}{u(f)/\left\{u(f)/v(f)\right\} + v(f)} = 3\frac{x}{v(f)}.$$

We now show that the following is less than or equal to 1:

$$\mathbb{E}\left[e^{\left|\underline{\mathcal{G}_n}(f)\right|/v(f)}\right] - 1 = \mathbb{E}\left[\int_0^{\left|\underline{\mathcal{G}_n}(f)\right|/v(f)} e^x dx\right] = \mathbb{E}\left[\int_0^{\infty} \mathbb{I}_{x \leq \left|\underline{\mathcal{G}_n}(f)\right|/v(f)} e^x dx\right]$$

Note the following:

- Since f is bounded and since $\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1} \leq \pi_{\min}^{T-1}$ a.s. by Condition 5.3.2, thus $\underline{\mathcal{G}_n}(f)$ is bounded a.s. (remember n is fixed).
- Since we are consider the cases in which $\|f\|_\infty > 0$, thus $v(f) > 0$.
- By the above two results, $\underline{\mathcal{G}_n}(f)/v(f)$ is bounded a.s., so $\mathbb{E}\left[\int_0^{\infty} \mathbb{I}_{x \leq \left|\underline{\mathcal{G}_n}(f)\right|/v(f)} e^x dx\right]$ is also bounded.

Thus, by Fubini's theorem, we can exchange the integrals,

$$= \int_0^{\infty} \mathbb{E}\left[\mathbb{I}_{x \leq \left|\underline{\mathcal{G}_n}(f)\right|/v(f)}\right] e^x dx = \int_0^{\infty} \mathbb{P}\Big(\left|\underline{\mathcal{G}_n}(f)\right| \geq xv(f)\Big) e^x dx$$

By display (C.7.19),

$$\leq 2\int_0^{\infty} e^{-3x} e^x dx = 2\int_0^{\infty} e^{-2x} dx = 2\left(\lim_{x\to\infty} -\frac{1}{2}e^{-2x} + \frac{1}{2}e^0\right) = 2\left(0 + \frac{1}{2}\right) = 1.$$

Thus we have that for $\gamma(x) = e^x - 1$,

$$\mathbb{E}\left[\gamma\left\{\left|\underline{\mathcal{G}_n}(f)\right|/v(f)\right\}\right] = \mathbb{E}\left[\exp\left\{\left|\underline{\mathcal{G}_n}(f)\right|/v(f)\right\}\right] - 1 \leq 1. \tag{C.7.20}$$

Note that $\gamma(x) = e^x - 1$ is convex, so by Jensen's inequality,

$$\exp\left(\mathbb{E}\left[\max_{f\in\mathcal{F}}\left|\underline{\mathcal{G}_n}(f)\right|/v(f)\right]\right) - 1 = \gamma\left(\mathbb{E}\left[\max_{f\in\mathcal{F}}\left|\underline{\mathcal{G}_n}(f)\right|/v(f)\right]\right)$$

$$\leq \mathbb{E}\left[\gamma\left(\max_{f\in\mathcal{F}}\left|\underline{\mathcal{G}_n}(f)\right|/v(f)\right)\right] \leq \sum_{f\in\mathcal{F}}\mathbb{E}\left[\gamma\left\{\left|\underline{\mathcal{G}_n}(f)\right|/v(f)\right\}\right] \leq |\mathcal{F}|.$$

The last inequality above holds by display (C.7.20). By adding 1 and taking the log on both sides, we have display (C.7.17) holds, i.e., that

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}\left|\underline{\mathcal{G}_n}(f)\right|/v(f)\right] \leq \log\left(|\mathcal{F}| + 1\right).$$

Proving display (C.7.18) ***holds.*** Let $x > 0$. We now state some results and discuss why they hold below. For all $n \geq 1$ we have that for any $f \in \mathcal{F}$,

$$\mathbb{P}\left(\left|\overline{\mathcal{G}}_n(f)\right| \geq x\right) \underbrace{=}_{(a)} \begin{cases} \mathbb{P}\left(\left|\overline{\mathcal{G}}_n(f)\right| \geq x\right) & \text{if } x < u(f)/v(f) \\ 0 & \text{if } x \geq u(f)/v(f) \end{cases} \tag{C.7.21}$$

$$\underbrace{\leq}_{(b)} \begin{cases} 2\exp\left(-6\frac{x^2}{u(f)+xv(f)}\right) & \text{if } x < u(f)/v(f) \\ 0 & \text{if } x \geq u(f)/v(f) \end{cases}$$

$$\underbrace{\leq}_{(c)} \begin{cases} 2\exp\left(-3\frac{x^2}{u(f)}\right) & \text{if } x < u(f)/v(f) \\ 0 & \text{if } x \geq u(f)/v(f) \end{cases} \leq 2\exp\left(-3\frac{x^2}{u(f)}\right).$$

- Inequality (a) holds because recall that $\overline{\mathcal{G}_n}(f) \triangleq |\mathbb{G}_n(f)|\mathbb{I}_{|\mathbb{G}_n(f)| \leq u(f)/v(f)}$.
- Inequality (b) holds by Lemma C.7.2 (Weighted Martingale Bernstein Inequality) since Condition 5.3.2 holds. Recall that $u(f) = 24\pi_{\min}^{-(T-1)}\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1} f(\mathcal{H}_T^{(i)})^2\right]$ and $v(f) = 24\pi_{\min}^{-(T-1)}\|f\|_\infty/\sqrt{n}$.
- Inequality (c) holds because if $x < u(f)/v(f)$, then

$$6\frac{x^2}{u(f) + xv(f)} \geq 6\frac{x^2}{u(f) + u(f)/v(f)v(f)} = 3\frac{x^2}{u(f)}.$$

We now show that the following is less than or equal to 1:

$$\mathbb{E}\left[e^{\left|\overline{\mathcal{G}_n}(f)\right|^2/u(f)}\right] - 1 = \mathbb{E}\left[\int_0^{\left|\overline{\mathcal{G}_n}(f)\right|^2/u(f)} e^x dx\right] = \mathbb{E}\left[\int_0^\infty \mathbb{I}_{x \leq \left|\overline{\mathcal{G}_n}(f)\right|^2/u(f)} e^x dx\right]$$

$$= \mathbb{E}\left[\int_0^\infty \mathbb{I}_{\sqrt{xu(f)} \leq \left|\overline{\mathcal{G}_n}(f)\right|} e^x dx\right]$$

Note the following:

- Since f is bounded and since $\left\{\hat{\pi}^{(i)}_{2:T}\right\}^{-1} \leq \pi_{\min}^{T-1}$ a.s. by Condition 5.3.2, thus $\underline{\mathcal{G}_n}(f)$ is bounded a.s. (remember n is fixed).
- Since are considering the cases in which $\mathbb{E}_{\pi^*_{2:t}}\left[\left\{\pi^{*,(i)}_{2:t}\right\}^{-1} f(\mathcal{H}_T^{(i)})^2\right] > 0$ (see discussion of special cases at the beginning of this proof), thus $u(f) > 0$.

- By the above two results, $\underline{\mathcal{G}_n}(f)/\sqrt{u(f)}$ is bounded a.s., so $\mathbb{E}\left[\int_0^\infty \mathbb{I}_{\sqrt{xu(f)}\leq|\overline{\mathcal{G}_n}(f)|}e^x dx\right]$ is also bounded.

Thus, by Fubini's theorem, we can exchange integrals,

$$= \int_0^\infty \mathbb{E}\left[\mathbb{I}_{\sqrt{xu(f)}\leq|\overline{\mathcal{G}_n}(f)|}\right] e^x dx = \int_0^\infty \mathbb{P}\left(\left|\overline{\mathcal{G}_n}(f)\right| \geq \sqrt{xu(f)}\right) e^x dx$$

By display (C.7.21),

$$\leq 2\int_0^\infty e^{-3x+x}dx = 2\int_0^\infty e^{-2x}dx = 2\left(\lim_{x\to\infty} -\frac{1}{2}e^{-2x} + \frac{1}{2}e^0\right) = 2\left(0+\frac{1}{2}\right) = 1.$$

Thus we have that for $\gamma_2(x) = e^{x^2} - 1$,

$$\gamma_2\left(\left|\overline{\mathcal{G}_n}(f)\right|/\sqrt{u(f)}\right) = \mathbb{E}\left[\exp\left(\left|\overline{\mathcal{G}_n}(f)\right|^2/u(f)\right)\right] - 1 \leq 1. \tag{C.7.22}$$

Since $\gamma_2(x) = e^{x^2} - 1$ is convex, by Jensen's inequality,

$$\exp\left(\mathbb{E}\left[\max_{f\in\mathcal{F}}\left|\overline{\mathcal{G}_n}(f)\right|/\sqrt{u(f)}\right]^2\right) - 1 = \gamma_2\left(\mathbb{E}\left[\max_{f\in\mathcal{F}}\left|\overline{\mathcal{G}_n}(f)\right|/\sqrt{u(f)}\right]\right)$$

$$\leq \mathbb{E}\left[\gamma_2\left(\max_{f\in\mathcal{F}}\left|\overline{\mathcal{G}_n}(f)\right|/\sqrt{u(f)}\right)\right] \leq \sum_{f\in\mathcal{F}}\mathbb{E}\left[\gamma_2\left(\left|\overline{\mathcal{G}_n}(f)\right|/\sqrt{u(f)}\right)\right] \leq |\mathcal{F}|.$$

The last inequality above holds by display (C.7.22). By adding 1, taking the log and the square-root on both sides, we have display (C.7.18) holds, i.e., that

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}\left|\overline{\mathcal{G}_n}(f)\right|/\sqrt{u(f)}\right] \leq \sqrt{\log\left(|\mathcal{F}|+1\right)}. \quad \blacksquare$$

C.8 Maximal Inequality as a Function of the Bracketing Integral (Lemma C.8.1)

Lemma C.8.1 (Maximal Inequality as a Function of the Bracketing Integral)**.** *Let $\delta > 0$. Let $\mathcal{F}$ be a class of real-valued measurable functions of $\mathcal{H}_t^{(i)}$ such that $\int_0^1 \sqrt{\log N_{[\,]}\big(\varepsilon, \mathcal{F}, L_2(\mathcal{P}_{\pi^*})\big)}d\varepsilon < \infty$ and $\mathbb{E}_{\pi^*_{2:T}}\big[\big\{\pi^{*,(i)}_{2:T}\big\}^{-1} f(\mathcal{H}_T^{(i)})^2\big] \leq \delta^2$ for all $f \in \mathcal{F}$. Under Condition 5.3.2 (Minimum Exploration), for all $n \geq 1$,*

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{F}} \left|\mathcal{G}_n(f)\right|\right] \lesssim \int_0^{\delta} \sqrt{\log N_{[\,]}\big(\delta, \mathcal{F}, L_2\big(\mathcal{P}_{\pi^*}\big)\big)}d\varepsilon + \sqrt{n}\mathbb{E}_{\pi^*_{2:t}}\left[\big\{\pi^{*,(i)}_{2:t}\big\}^{-1} F(\mathcal{H}_t^{(i)})\mathbb{I}_{F(\mathcal{H}_t^{(i)})>\sqrt{n}a(\delta)}\right], \quad \text{(C.8.1)}$$

where

- $a(\delta) = \delta/\sqrt{\log N_{[\,]}\big(\delta, \mathcal{F}, L_2(\mathcal{P}_{\pi^*})\big)}$
- $\mathcal{G}_n(f) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\big\{\hat{\pi}^{(i)}_{2:t}\big\}^{-1} f(\mathcal{H}_t^{(i)}) - \mathbb{E}\left[\big\{\hat{\pi}^{(i)}_{2:t}\big\}^{-1} f(\mathcal{H}_t^{(i)})\right]\right)$
- *F is an envelope where $\sup_{f\in\mathcal{F}} |f(\mathcal{H}_t^{(i)})| < F(\mathcal{H}_t^{(i)}) < \infty$ with probability 1.*

Above $\lesssim$ means less than or equal to when scaled by universal positive constants. Above $\mathbb{E}^*$ refers to outer expectations as defined in Section 18.2[99]*.*

Proof of Lemma C.8.1 (Maximal Inequality as a Function of the Bracketing Integral). Our proof is almost identical to that of Lemma 19.34[99] except that we use the maximal inequality in Lemma C.7.3 instead of a maximal inequality for i.i.d. data; we include the full proof for clarity and completeness. For notational convenience we consider the t set to T case; the argument holds by the same argument for any $t \in [2\colon T]$.

Note that by triangle inequality,

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{F}}|\mathcal{G}_n(f)|\right] \leq \mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\mathcal{G}_n\left(f\mathbb{I}_{F>\sqrt{n}a(\delta)}\right)\right|\right] + \mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\mathcal{G}_n\left(f\mathbb{I}_{F\leq\sqrt{n}a(\delta)}\right)\right|\right]. \quad \text{(C.8.2)}$$

Bounding First Term in display (C.8.2)**.** This term is to deal with potentially unbounded functions $f \in \mathcal{F}$.

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\mathcal{G}_n\left(f\mathbb{I}_{F>\sqrt{n}a(\delta)}\right)\right|\right]$$

By using the definition of $\mathcal{G}_n$,

$$= \mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)})\mathbb{I}_{F(\mathcal{H}_T^{(i)})>\sqrt{n}a(\delta)} - \mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)})\mathbb{I}_{F(\mathcal{H}_T^{(i)})>\sqrt{n}a(\delta)}\right]\right)\right|\right]$$

By triangle inequality,

$$\leq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)})\mathbb{I}_{F(\mathcal{H}_T^{(i)})>\sqrt{n}a(\delta)}\right|\right] + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sup_{f\in\mathcal{F}}\left\{\left|\mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}f(\mathcal{H}_T^{(i)})\mathbb{I}_{F(\mathcal{H}_T^{(i)})>\sqrt{n}a(\delta)}\right]\right|\right\}$$

By Jensen's inequality,

$$\leq 2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left(\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}\left|f(\mathcal{H}_T^{(i)})\right|\mathbb{I}_{F(\mathcal{H}_T^{(i)})>\sqrt{n}a(\delta)}\right)\right]$$

Recall our envelope function F satisfies $|f(\mathcal{H}_t^{(i)})| < F(\mathcal{H}_t^{(i)}) < \infty$ a.s., so

$$\leq 2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\left[\left\{\hat{\pi}_{2:T}^{(i)}\right\}^{-1}F(\mathcal{H}_T^{(i)})\mathbb{I}_{F(\mathcal{H}_T^{(i)})>\sqrt{n}a(\delta)}\right]$$

$$= 2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}_{\pi_{2:T}^*}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1}F(\mathcal{H}_T^{(i)})\mathbb{I}_{F(\mathcal{H}_T^{(i)})>\sqrt{n}a(\delta)}\right]$$

Since the expectation above is indexed by the deterministic policy $\pi_{2:T}^*$, $\mathcal{H}_T^{(i)}$ within the expectation are i.i.d.

$$= 2\sqrt{n}\mathbb{E}_{\pi_{2:T}^*}\left[\left\{\pi_{2:T}^{*,(i)}\right\}^{-1}F(\mathcal{H}_T^{(i)})\mathbb{I}_{F(\mathcal{H}_T^{(i)})>\sqrt{n}a(\delta)}\right]$$

This us gives us the second part of the bound from display (C.8.1).

Bounding Second Term in display **(C.8.2).** We now focus on bounding the following:

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\mathcal{G}_n f\mathbb{I}_{F\leq\sqrt{n}a(\delta)}\right|\right]$$

We now consider the class of functions $\bar{\mathcal{F}} := \left\{f\mathbb{I}_{F\leq\sqrt{n}a(\delta)} : f\in\mathcal{F}\right\}$. We first show that $N_{[\,]}\left(\varepsilon,\bar{\mathcal{F}},L_2(\mathcal{P}_{\pi^*})\right) \leq N_{[\,]}\left(\varepsilon,\mathcal{F},L_2(\mathcal{P}_{\pi^*})\right)$.

- By definition of bracketing numbers, we can cover $\mathcal{F}$ with $N_{[\,]}\left(\varepsilon,\mathcal{F},L_2(\mathcal{P}_{\pi^*})\right)$ brackets, each with size at most ε. Specifically, we can find brackets $[l_j,u_j]$ for $j\in\left[1\colon N_{[\,]}\left(\varepsilon,\mathcal{F},L_2(\mathcal{P}_{\pi^*})\right)\right]$ that cover $\mathcal{F}$ such that $\mathbb{E}_{\pi_{2:T}^*}\left[(u_j-l_j)^2\right]^{1/2} \leq \varepsilon$ for all brackets $[l_j,u_j]$.

- Note that brackets $\left[l_j\mathbb{I}_{F\leq\sqrt{n}a(\delta)},\ u_j\mathbb{I}_{F\leq\sqrt{n}a(\delta)}\right]$ for some $j\in\left[1\colon N_{[\,]}\left(\varepsilon,\mathcal{F},L_2(\mathcal{P}_{\pi^*})\right)\right]$

cover $\bar{\mathcal{F}}$.

- Additionally, note that $\mathbb{E}_{\pi^*_{2:T}}\left[\left([u_j - l_j]\mathbb{I}_{F\leq\sqrt{n}a(\delta)}\right)^2\right] \leq \mathbb{E}_{\pi^*_{2:T}}\left[(u_j - l_j)^2\right] \leq \varepsilon^2$.

Thus, we have that

$$N_{[\,]}\big(\varepsilon, \bar{\mathcal{F}}, L_2(\mathcal{P}_{\pi^*})\big) \leq N_{[\,]}\big(\varepsilon, \mathcal{F}, L_2(\mathcal{P}_{\pi^*})\big). \tag{C.8.3}$$

Desiderata for Nested Partitions. We now assume the existence of nested partitions of $\bar{\mathcal{F}}$ that satisfy certain conditions. We will finish the proof assuming these partitions exist and conclude by constructing these partitions.

High level, we assume we have nested partitions of $\bar{\mathcal{F}}$ that are indexed by positive integers q. These partitions are designed to become increasingly fine-grained as q increases. Specifically the "size" of each piece of the partition will be on the order of 2^{-q}, i.e., the "size" of the partitions will halve as q increases by 1. The partitions are nested in that each partition piece at level $q+1$ is a subset of some partition piece at level q.

We pick q_0 to be a positive integer such that

$$\delta < 2^{-q_0} \leq 2\delta. \tag{C.8.4}$$

For every integer $q \geq q_0$ we have a partition of $\bar{\mathcal{F}}$; we denote these partions as $\{\bar{\mathcal{F}}_{q,j}\}_{j=1}^{N_q}$. We assume that $N_{q_0} = N_{[\,]}\big(2^{-q_0}, \bar{\mathcal{F}}, L_2(\mathcal{P}_{\pi^*})\big)$. These partitions are nested in that for each $q \geq q_0 + 1$ and for every $j \in [1\colon N_q]$, we have that the partition piece $\bar{\mathcal{F}}_{q,j}$ is a subset of some partition piece $\bar{\mathcal{F}}_{q-1,k}$ for some $k \in [1\colon N_{q-1}]$. Moreover, we further assume the following:

- **Requirement on the "size" of partition pieces:** For each partition q and partition

piece $j \in [1\colon N_q]$, let $\Delta_{q,j}$ be a measurable function of $\mathcal{H}_T^{(i)}$ such that $\sup_{f,g\in\bar{\mathcal{F}}_{q,j}} |f - g| \le \Delta_{q,j}$ a.s. and

$$\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\Delta_{q,j}(\mathcal{H}_T^{(i)})^2\right] \le 2^{-2q}. \tag{C.8.5}$$

- **Requirement on how the number of partition pieces grows as the "size" goes to zero:**

$$\sum_{q=q_0}^{\infty} 2^{-q}\sqrt{\log N_q} \lesssim \int_0^{\delta} \sqrt{\log N_{[\,]}\big(\varepsilon, \bar{\mathcal{F}}, L_2(\mathcal{P}_{\pi^*})\big)}d\varepsilon. \tag{C.8.6}$$

We construct nested partitions that satisfy the above conditions at the end of this proof. For now, we assume such nested partitions described above exist and we continue with the argument.

Main Argument Assuming Desired Nested Partitions Exist. For every partition piece $\bar{\mathcal{F}}_{q,j}$, we choose a arbitrary point $\bar{f}_{q,j}$ in that partition piece, i.e., for each $q \ge q_0$ and every $j \in [1\colon N_q]$ we choose a point $\bar{f}_{q,j} \in \mathcal{F}_{q,j}$. We also define functions $\lambda_q : \bar{\mathcal{F}} \mapsto \bar{\mathcal{F}}$ that maps each function $\bar{f} \in \bar{\mathcal{F}}$ to these points $\{\bar{f}_{q,j}\}_{j=1}^{N_q}$; specifically, for any $\bar{f} \in \bar{\mathcal{F}}$ we can find some partition piece $\bar{\mathcal{F}}_{q,j}$ such that $\bar{f} \in \bar{\mathcal{F}}_{q,j}$ and we map that $\bar{f}$ to the point $\bar{f}_{q,j}$.
Note that for any integer $Q > q_0$, by telescoping series, for any $\bar{f} \in \bar{\mathcal{F}}$,

$$\bar{f}(\mathcal{H}_T^{(i)}) = \lambda_{q_0}\bar{f}(\mathcal{H}_T^{(i)}) + \sum_{q=q_0}^{Q}\left\{\lambda_{q+1}\bar{f}(\mathcal{H}_T^{(i)}) - \lambda_q\bar{f}(\mathcal{H}_T^{(i)})\right\} + \bar{f}(\mathcal{H}_T^{(i)}) - \lambda_{Q+1}\bar{f}(\mathcal{H}_T^{(i)})$$

$$= \lambda_{q_0}\bar{f}(\mathcal{H}_T^{(i)}) + \sum_{q=q_0}^{\infty} \mathbb{I}_{q\le Q}\left\{\lambda_{q+1}\bar{f}(\mathcal{H}_T^{(i)}) - \lambda_q\bar{f}(\mathcal{H}_T^{(i)})\right\} + \sum_{q=q_0}^{\infty} \mathbb{I}_{q=Q+1}\left\{\bar{f}(\mathcal{H}_T^{(i)}) - \lambda_q\bar{f}(\mathcal{H}_T^{(i)})\right\}. \quad \text{(C.8.7)}$$

For any $\bar{f} \in \bar{\mathcal{F}}$, we define $Q_{\bar{f}}(\mathcal{H}_T^{(i)}) \in [q_0, \infty]$ to be a random variable representing the maximum partition level with no bound violations up to that level. Specifically,

$$Q_{\bar{f}}(\mathcal{H}_T^{(i)}) \triangleq \left\{ \sup_{q\ge q_0} \text{ s.t. } \sum_{j=1}^{N_p} \mathbb{I}_{\bar{f}\in\bar{\mathcal{F}}_{p,j}} \Delta_{p,j}(\mathcal{H}_T^{(i)}) \le \sqrt{n}2^{-p}/\sqrt{\log N_p} \text{ for all } p \in [q_0 : q] \right\}, \quad \text{(C.8.8)}$$

Thus, by replacing Q with $Q_{\bar{f}}$ and by applying $\mathcal{G}_n$ to both sides of display (C.8.7),

$$\mathcal{G}_n(\bar{f}) = \mathcal{G}_n(\lambda_{q_0}\bar{f}) + \sum_{q=q_0}^{\infty} \mathcal{G}_n\left(\mathbb{I}_{q\le Q_{\bar{f}}}\left(\lambda_{q+1}\bar{f} - \lambda_q\bar{f}\right)\right) + \sum_{q=q_0}^{\infty} \mathcal{G}_n\left(\mathbb{I}_{q=Q_{\bar{f}}+1}\left(\bar{f} - \lambda_q\bar{f}\right)\right)$$

Thus, we have that by triangle inequality

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{F}} \left|\mathcal{G}_n f \mathbb{I}_{F\le\sqrt{n}a(\vartheta)}\right|\right] = \mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}} \left|\mathcal{G}_n(\bar{f})\right|\right]$$
$$\le \underbrace{\mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}} \left|\mathcal{G}_n(\lambda_{q_0}\bar{f})\right|\right]}_{(i)} + \underbrace{\mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}} \left|\sum_{q=q_0}^{\infty} \mathcal{G}_n\left(\mathbb{I}_{q\le Q_{\bar{f}}}\left(\lambda_{q+1}\bar{f} - \lambda_q\bar{f}\right)\right)\right|\right]}_{(ii)}$$
$$+ \underbrace{\mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}} \left|\sum_{q=q_0}^{\infty} \mathcal{G}_n\left(\mathbb{I}_{q=Q_{\bar{f}}+1}\left(\bar{f} - \lambda_q\bar{f}\right)\right)\right|\right]}_{(iii)}. \quad \text{(C.8.9)}$$

Below we will show the following results:

- Bounding term (i)

$$\mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\mathcal{G}_n(\lambda_{q_0}\bar{f})\right|\right] \lesssim 2\pi_{\min}^{-(T-1)}2^{-q_0}\sqrt{\log N_{q_0}} \tag{C.8.10}$$

- Bounding term (ii)

$$\mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\sum_{q=q_0}^{\infty}\mathcal{G}_n\left(\mathbb{I}_{q\le Q_{\bar{f}}}\left(\lambda_{q+1}\bar{f}-\lambda_q\bar{f}\right)\right)\right|\right] \lesssim 2\pi_{\min}^{-(T-1)}\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log N_q} \tag{C.8.11}$$

- Bounding term (iii)

$$\mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\sum_{q=q_0}^{\infty}\mathcal{G}_n\left(\mathbb{I}_{q=Q_{\bar{f}}+1}\left(\bar{f}-\lambda_q\bar{f}\right)\right)\right|\right] \lesssim 4\pi_{\min}^{-(T-1)}\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log N_q}. \tag{C.8.12}$$

For now we assume the above three displays hold (we show they hold later). Thus, we can upper bound display (C.8.9) as follows:

$$\mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\mathcal{G}_n(\bar{f})\right|\right] \lesssim 2\pi_{\min}^{-(T-1)}2^{-q_0}\sqrt{\log N_{q_0}} + 6\pi_{\min}^{-(T-1)}\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log N_q}$$

$$\le 8\pi_{\min}^{-(T-1)}\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log N_q}$$

By display (C.8.6),

$$\lesssim \pi_{\min}^{-(T-1)}\int_0^{\delta}\sqrt{\log(N_{[\,]}\left(\varepsilon,\bar{\mathcal{F}},L_2(\mathcal{P}_{\pi^*})\right)}d\varepsilon \le \pi_{\min}^{-(T-1)}\int_0^{\delta}\sqrt{\log(N_{[\,]}\left(\varepsilon,\mathcal{F},L_2(\mathcal{P}_{\pi^*})\right)}d\varepsilon.$$

The last inequality above holds by display (C.8.3). We now show that displays (C.8.10),

(C.8.11), and (C.8.12) hold.

Display (C.8.10)*: Bounding term (i)*.

$$\mathbb{E}^*\left[\sup_{\bar{f}\in\bar{\mathcal{F}}}\left|\mathcal{G}_n(\lambda_{q_0}\bar{f})\right|\right] = \mathbb{E}\left[\max_{j\in[1:\,N_{q_0}]}\left|\mathcal{G}_n(\bar{f}_{q_0,j})\right|\right]$$

By Lemma C.7.3 (Maximal Inequality for Finite Class of Functions) for any $n \geq 1$,

$$\lesssim \pi_{\min}^{-(T-1)} \max_{j\in[1:\,N_{q_0}]} \frac{\|\bar{f}_{q_0,j}\|_\infty}{\sqrt{n}} \log N_{q_0} + \sqrt{\pi_{\min}^{-(T-1)}} \max_{j\in[1:\,N_{q_0}]} \sqrt{\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\bar{f}_{q_0,j}(\mathcal{H}_T^{(i)})^2\right]}\sqrt{\log N_{q_0}}. \quad \text{(C.8.13)}$$

- Note that since $\bar{f}_{q_0,j}(\mathcal{H}_T^{(i)}) = f_{q_0,j}(\mathcal{H}_T^{(i)})\mathbb{I}_{F(\mathcal{H}_T^{(i)})\leq\sqrt{n}a(\delta)} \leq \sqrt{n}a(\delta)$ a.s., we get the first inequality below:

$$\max_{j\in[1:\,N_{q_0}]}\left\{\|\bar{f}_{q_0,j}\|_\infty\right\} \leq \sqrt{n}a(\delta) = \sqrt{n}\delta/\sqrt{\log N_{[\,]}\big(\delta,\mathcal{F},L_2(\mathcal{P}_{\pi^*})\big)} \leq \sqrt{n}2^{-q_0}/\sqrt{\log N_{q_0}}.$$

  The last inequality above holds because $\delta < 2^{-q_0} \leq 2\delta$ from display (C.8.4) and

$$N_{[\,]}\big(\delta,\mathcal{F},L_2(\mathcal{P}_{\pi^*})\big) \geq N_{[\,]}\big(2^{-q_0},\mathcal{F},L_2(\mathcal{P}_{\pi^*})\big) \geq N_{[\,]}\big(2^{-q_0},\bar{\mathcal{F}},L_2(\mathcal{P}_{\pi^*})\big) = N_{q_0}.$$

  The last inequality above holds by display (C.8.3).

- $\max_{j\in[1:\,N_q]}\sqrt{\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\bar{f}_{q_0,j}(\mathcal{H}_T^{(i)})^2\right]} \leq \delta \leq 2^{-q_0}$; the first inequality holds by assumption of the Lemma and the second inequality holds since we choose q_0

such that $\delta < 2^{-q_0} \leq 2\delta$.

Thus, by the above bullets, we can upper bound display (C.8.13) as follows:

$$\leq \pi_{\min}^{-(T-1)} 2^{-q_0} \sqrt{\log N_{q_0}} + \sqrt{\pi_{\min}^{-(T-1)}} 2^{-q_0} \sqrt{\log N_{q_0}} \leq 2\pi_{\min}^{-(T-1)} 2^{-q_0} \sqrt{\log N_{q_0}}.$$

Thus, we have that display (C.8.10) holds.

Bounding term (ii):. By triangle inequality,

$$\mathbb{E}^* \left[\sup_{\bar{f} \in \bar{\mathcal{F}}} \left| \sum_{q=q_0}^{\infty} \mathcal{G}_n \left(\mathbb{I}_{q \leq Q_{\bar{f}}} \left(\lambda_{q+1} \bar{f} - \lambda_q \bar{f} \right) \right) \right| \right] \leq \sum_{q=q_0}^{\infty} \mathbb{E}^* \left[\sup_{\bar{f} \in \bar{\mathcal{F}}} \left| \mathcal{G}_n \left(\mathbb{I}_{q \leq Q_{\bar{f}}} \left(\lambda_{q+1} \bar{f} - \lambda_q \bar{f} \right) \right) \right| \right]$$

By the definition of $Q_{\bar{f}}$ from display (C.8.8), $\mathbb{I}_{q \leq Q_{\bar{f}}} = \mathbb{I}_{q \leq Q_{\lambda_q \bar{f}}}$,

$$= \sum_{q=q_0}^{\infty} \mathbb{E} \left[\max_{j \in [1 \colon N_q]} \left| \mathcal{G}_n \left(\mathbb{I}_{q \leq Q_{\bar{f}_{q,j}}} \left(\bar{f}_{q+1,j} - \bar{f}_{q,j} \right) \right) \right| \right]$$

By Lemma C.7.3 (Maximal Inequality for Finite Class of Functions) for all sufficiently large n,

$$\lesssim \sum_{q=q_0}^{\infty} \pi_{\min}^{-(T-1)} \max_{j \in [1 \colon N_{q_0}]} \frac{\left\| \mathbb{I}_{q \leq Q_{\bar{f}_{q,j}}} \left(\bar{f}_{q+1,j} - \bar{f}_{q,j} \right) \right\|_{\infty}}{\sqrt{n}} \log N_q$$
$$+ \sum_{q=q_0}^{\infty} \sqrt{\pi_{\min}^{-(T-1)}} \max_{j \in [1 \colon N_q]} \sqrt{\mathbb{E}_{\pi_{2:T}^*} \left[\left\{ \pi_{2:T}^{*,(i)} \right\}^{-1} \mathbb{I}_{q \leq Q_{\bar{f}_{q,j}}} \left(\bar{f}_{q+1,j} - \bar{f}_{q,j} \right)^2 \right]} \sqrt{\log N_q}. \quad \text{(C.8.14)}$$

- Note that by the definition of $\mathbb{I}_{q \leq Q_{\bar{f}_{q,j}}}$ ($Q_{\bar{f}}$ was defined in display (C.8.8)) and by our

nested partitions, we have that

$$\left\|\mathbb{I}_{q\le Q_{\bar{f}_{q,j}}}\left(\bar{f}_{q+1,j}-\bar{f}_{q,j}\right)\right\|_\infty \le \sup_{\bar{f},\bar{f}'\in\bar{\mathcal{F}}_{q,j}}\left\|\mathbb{I}_{q\le Q_{\bar{f}_{q,j}}}|f-f'|\right\|_\infty$$
$$\le \left\|\mathbb{I}_{q\le Q_{\bar{f}_{q,j}}}\Delta_{q,j}(\mathcal{H}_T^{(i)})\right\|_\infty \le \sqrt{n}2^{-q}/\sqrt{\log N_q}.$$

- By our nested partitions, we have that $\lambda_{q+1}\bar{f}, \lambda_q\bar{f}$ are in the same q^{th}-level partition piece, i.e., $\lambda_{q+1}\bar{f}, \lambda_q\bar{f} \in \bar{\mathcal{F}}_{q,j}$ for some $\bar{\mathcal{F}}_{q,j}$ with $j \in [1\colon N_q]$. Thus,

$$\mathbb{E}_{\pi^*_{2:T}}\left[\mathbb{I}_{q\le Q_{\bar{f}_{q,j}}(\mathcal{H}_T^{(i)})}\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\left(\bar{f}_{q+1,j}(\mathcal{H}_T^{(i)})-\bar{f}_{q,j}(\mathcal{H}_T^{(i)})\right)^2\right]$$
$$\le \sup_{\bar{f},\bar{f}'\in\bar{\mathcal{F}}_{q,j}}\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\left(\bar{f}(\mathcal{H}_T^{(i)})-\bar{f}'(\mathcal{H}_T^{(i)})\right)^2\right]$$
$$\le \mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\Delta_{q,j}(\mathcal{H}_T^{(i)})^2\right].$$

Moreover, by properties of our partitions,

$$\max_{j\in[1\colon N_q]}\sqrt{\mathbb{E}_{\pi^*_{2:T}}\left[\mathbb{I}_{q\le Q_{\bar{f}_{q,j}}(\mathcal{H}_T^{(i)})}\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\left(\bar{f}_{q+1,j}(\mathcal{H}_T^{(i)})-\bar{f}_{q,j}(\mathcal{H}_T^{(i)})\right)^2\right]}$$
$$\le \max_{j\in[1\colon N_q]}\sqrt{\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\Delta_{q,j}(\mathcal{H}_T^{(i)})^2\right]} \le 2^{-q}.$$

The last inequality above holds by the size property of our partitions from display (C.8.5).

By the above bullets, we have that display (C.8.14) is upper bounded by the following:

$$\leq \sum_{q=q_0}^{\infty} \left\{ \pi_{\min}^{-(T-1)} 2^{-q} \sqrt{\log N_q} + \sqrt{\pi_{\min}^{-(T-1)}} 2^{-q} \sqrt{\log N_q} \right\} \leq 2\pi_{\min}^{-(T-1)} \sum_{q=q_0}^{\infty} 2^{-q} \sqrt{\log N_q}.$$

Thus, we have that display (C.8.11) holds.

Bounding term (iii):. By triangle inequality,

$$\mathbb{E}^* \left[\sup_{\bar{f} \in \bar{\mathcal{F}}} \left| \sum_{q=q_0}^{\infty} \mathcal{G}_n \left(\mathbb{I}_{q=Q_{\bar{f}}+1} \left(\bar{f} - \lambda_q \bar{f} \right) \right) \right| \right] \leq \sum_{q=q_0}^{\infty} \mathbb{E}^* \left[\sup_{\bar{f} \in \bar{\mathcal{F}}} \left| \mathcal{G}_n \left(\mathbb{I}_{q=Q_{\bar{f}}+1} \left(\bar{f} - \lambda_q \bar{f} \right) \right) \right| \right] \tag{C.8.15}$$

Note that if some functions f, g are such that $\left| f(\mathcal{H}_T^{(i)}) \right| \leq g(\mathcal{H}_T^{(i)})$ a.s., then

$$\begin{aligned} |\mathcal{G}_n(f)| &\leq \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) \right| + \sqrt{n} \left| \mathbb{E} \left[\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} f(\mathcal{H}_T^{(i)}) \right] \right| \\ &\leq \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} g(\mathcal{H}_T^{(i)}) \right| + \sqrt{n} \left| \mathbb{E} \left[\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} g(\mathcal{H}_T^{(i)}) \right] \right| \\ &\leq |\mathcal{G}_n(g)| + 2\sqrt{n} \left| \mathbb{E} \left[\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} g(\mathcal{H}_T^{(i)}) \right] \right| \text{ a.s.} \end{aligned} \tag{C.8.16}$$

Note that $\left| \mathbb{I}_{q=Q_{\bar{f}}(\mathcal{H}_T^{(i)})+1} \left(\bar{f}(\mathcal{H}_T^{(i)}) - \lambda_q \bar{f}(\mathcal{H}_T^{(i)}) \right) \right| \leq \mathbb{I}_{q=Q_{\bar{f}}(\mathcal{H}_T^{(i)})+1} \sum_{j=1}^{N_q} \mathbb{I}_{\bar{f} \in \bar{\mathcal{F}}_{q,j}} \Delta_{q,j}(\mathcal{H}_T^{(i)})$ a.s. by the definition of $\Delta_{q,j}$ from above display (C.8.5). So by the observation from display

(C.8.16) we can upper bound display (C.8.15) as follows:

$$\leq \sum_{q=q_0}^{\infty} \mathbb{E}^* \left[\sup_{\bar{f} \in \bar{\mathcal{F}}} \left| \mathcal{G}_n \left(\mathbb{I}_{q=Q_{\bar{f}}(\mathcal{H}_T^{(i)})+1} \sum_{j=1}^{N_q} \mathbb{I}_{\bar{f} \in \mathcal{F}_{q,j}} \Delta_{q,j}(\mathcal{H}_T^{(i)}) \right) \right| \right]$$
$$+ 2\sqrt{n} \sum_{q=q_0}^{\infty} \sup_{\bar{f} \in \bar{\mathcal{F}}} \left| \mathbb{E} \left[\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} \mathbb{I}_{q=Q_{\bar{f}}(\mathcal{H}_T^{(i)})+1} \sum_{j=1}^{N_q} \mathbb{I}_{\bar{f} \in \mathcal{F}_{q,j}} \Delta_{q,j}(\mathcal{H}_T^{(i)}) \right] \right|$$

Since $\mathbb{I}_{q \leq Q_{\bar{f}}+1} = \mathbb{I}_{q \leq Q_{\lambda_q \bar{f}}+1}$ ($Q_{\bar{f}}$ was defined in display (C.8.8)),

$$= \sum_{q=q_0}^{\infty} \mathbb{E} \left[\max_{j \in [1:\, N_q]} \left| \mathcal{G}_n \left(\mathbb{I}_{q=Q_{\bar{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} \, \Delta_{q,j}(\mathcal{H}_T^{(i)}) \right) \right| \right]$$
$$+ 2\sqrt{n} \sum_{q=q_0}^{\infty} \max_{j \in [1:\, N_q]} \mathbb{E} \left[\left\{ \hat{\pi}_{2:t}^{(i)} \right\}^{-1} \mathbb{I}_{q=Q_{\bar{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} \Delta_{q,j}(\mathcal{H}_T^{(i)}) \right] \quad \text{(C.8.17)}$$

- Due to the nested property of our partitions (see above display (C.8.5) for the definition of $\Delta_{q,j}$),

$$\mathbb{I}_{q=Q_{\bar{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} \, \Delta_{q,j}(\mathcal{H}_T^{(i)}) \leq \mathbb{I}_{q=Q_{\bar{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} \, \Delta_{q-1,j}(\mathcal{H}_T^{(i)}) \text{ a.s.}$$

  Moreover, by the definition of $Q_{\bar{f}_{q,j}}$ from display (C.8.8),

$$\mathbb{I}_{q=Q_{\bar{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} \, \Delta_{q-1,j}(\mathcal{H}_T^{(i)}) \leq \sqrt{n} 2^{-(q-1)} / \sqrt{\log N_{q-1}} \text{ a.s.}$$

  Thus, by Lemma C.7.3 (Maximal Inequality for Finite Class of Functions) and dis-

play (C.8.5),

$$\sum_{q=q_0}^{\infty} \mathbb{E}^* \left[\max_{j \in [1:\, N_q]} \left| \mathcal{G}_n \left(\mathbb{I}_{q = Q_{\tilde{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} \, \Delta_{q,j}(\mathcal{H}_T^{(i)}) \right) \right| \right]$$
$$\lesssim \sum_{q=q_0}^{\infty} \left\{ \pi_{\min}^{-(T-1)} 2^{-(q-1)} / \sqrt{\log N_{q-1}} \log N_q + \sqrt{\pi_{\min}^{-(T-1)}} 2^{-q} \sqrt{\log N_q} \right\}$$

Since $N_{q-1} \leq N_q$ (bracketing number for brackets of size $2^{-(q-1)}$ vs 2^{-q}), thus $\log N_{q-1} \leq \log N_q$ and $1 \leq \sqrt{\log N_q / \log N_{q-1}}$. So,

$$\leq \sum_{q=q_0}^{\infty} \left\{ \pi_{\min}^{-(T-1)} 2^{-(q-1)} \sqrt{\log N_q} + \sqrt{\pi_{\min}^{-(T-1)}} 2^{-q} \sqrt{\log N_q} \right\}$$
$$\leq \pi_{\min}^{-(T-1)} 2 \sum_{q=q_0}^{\infty} 2^{-q} \sqrt{\log N_q}.$$

- Note that by our nested partitions and by the definition of $Q_{\tilde{f}_{q,j}}$ from display (C.8.8), if $\mathbb{I}_{q = Q_{\tilde{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} = 1$, then $\Delta_{q,j}(\mathcal{H}_T^{(i)}) > \sqrt{n} 2^{-q} / \sqrt{\log N_q}$. Thus, when $\mathbb{I}_{q = Q_{\tilde{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} = 1$, $\Delta_{q,j}(\mathcal{H}_T^{(i)}) \left(\sqrt{n} 2^{-q} / \sqrt{\log N_q} \right)^{-1} \geq 1$. Thus,

$$\mathbb{I}_{q = Q_{\tilde{f}_{q,j}}(\mathcal{H}_T^{(i)})+1} \Delta_{q,j}(\mathcal{H}_T^{(i)}) \leq \Delta_{q,j}(\mathcal{H}_T^{(i)})^2 \left(\sqrt{n} 2^{-q} / \sqrt{\log N_q} \right)^{-1}.$$

Thus we have that

$$2\sqrt{n}\sum_{q=q_0}^{\infty}\max_{j\in[1:\,N_q]}\left|\mathbb{E}\left[\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1}\mathbb{I}_{q=Q_{\tilde{f}_{q,j}}(\mathcal{H}^{(i)}_T)+1}\Delta_{q,j}(\mathcal{H}^{(i)}_T)\right]\right|$$
$$\leq\sum_{q=q_0}^{\infty}\frac{2\sqrt{\log N_q}}{2^{-q}}\max_{j\in[1:\,N_q]}\mathbb{E}\left[\left\{\hat{\pi}^{(i)}_{2:t}\right\}^{-1}\Delta_{q,j}(\mathcal{H}^{(i)}_T)^2\right]$$
$$\leq\sum_{q=q_0}^{\infty}\frac{2\sqrt{\log N_q}}{2^{-q}}\cdot 2^{-2q}=2\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log N_q}.$$

The last inequality above holds by the size property of our partitions from display (C.8.5).

By the observations in the above bullets, we can upper bound display (C.8.17) as follows:

$$\lesssim\pi_{\min}^{-(T-1)}2\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log N_q}+2\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log N_q}$$

$$\leq\pi_{\min}^{-(T-1)}4\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log N_q}$$

Thus, we have that display (C.8.12) holds.

Construct nested partitions. We now construct nested partitions that satisfy the conditions described previously, particularly displays (C.8.5) and (C.8.6).

By our bracketing number assumption, for every integer $q\geq q_0$, we can find $N^*_q\triangleq N_{[\,]}\big(2^{-q}\pi_{\min}^{(T-1)/2},\bar{\mathcal{F}},L_2(\mathcal{P}_{\pi^*})\big)$ bracketing functions $\left\{\left[l^*_{q,j},u^*_{q,j}\right]\right\}_{j=1}^{N^*_q}$ of size at most $2^{-q}\pi_{\min}^{(T-1)/2}$ that cover $\bar{\mathcal{F}}$, i.e.,

$$\mathbb{E}_{\pi^*_{2:T}}\left[\left(u^*_{q,j}(\mathcal{H}^{(i)}_T)-l^*_{q,j}(\mathcal{H}^{(i)}_T)\right)^2\right]^{1/2}\leq 2^{-q}\pi_{\min}^{(T-1)/2}. \tag{C.8.18}$$

These brackets form a partition of $\bar{\mathcal{F}}$, which we write as $\left\{\bar{\mathcal{F}}^*_{q,j}\right\}_{j=1}^{N^*_q}$. Note that these partitions are not necessarily nested. Below we use $\left\{\bar{\mathcal{F}}^*_{q,j}\right\}_{j=1}^{N^*_q}$ to refer to the potentially **non-nested** partitions (that exist by assumption) and $\left\{\bar{\mathcal{F}}_{q,j}\right\}_{j=1}^{N_q}$ to refer to the **nested** partitions (that we will construct).

We take intersections of the partitions $\left\{\bar{\mathcal{F}}^*_{q,j}\right\}_{j=1}^{N^*_q}$ to construct a set of nested partitions $\left\{\bar{\mathcal{F}}_{q,j}\right\}_{j=1}^{N_q}$ for all integers $q \geq q_0$.

- For partition $\left\{\bar{\mathcal{F}}_{q_0,j}\right\}_{j=1}^{N_{q_0}}$, we simply set $\bar{\mathcal{F}}_{q_0,j} \triangleq \bar{\mathcal{F}}^*_{q_0,j}$ for all $j \in \left[1\colon N^*_{q_0}\right]$. This means that $N_{q_0} \triangleq N^*_{q_0}$.

- For partition $\left\{\bar{\mathcal{F}}_{q_0+1,j}\right\}_{j=1}^{N_{q_0+1}}$, we set partition pieces $\bar{\mathcal{F}}_{q_0+1,j}$ for all $j \in \left[1\colon N^*_{q_0}\right]$ to be the intersections between all pairs of partition pieces $\bar{\mathcal{F}}^*_{q_0,k}$ and $\bar{\mathcal{F}}^*_{q_0+1,l}$ for $k \in \left[1\colon N_{q_0}\right]$ and $l \in \left[1\colon N_{q_0+1}\right]$. This means that $N_{q_0+1} \triangleq N^*_{q_0} \cdot N^*_{q_1}$. Note that it could be that some partition pieces $\bar{\mathcal{F}}_{q_0+1,j}$ are empty; this is okay, as we simply want to upper bound the bracketing number.

- For general $q \geq q_0$, we set partition pieces $\bar{\mathcal{F}}_{q,j}$ for $j \in \left[1\colon N_q\right]$ to be the intersections between all possible combinations in which we take one partition piece from each partition level $\left\{\bar{\mathcal{F}}^*_{p,j}\right\}_{j=1}^{N^*_p}$ for each $p \in [q_0\colon q]$. This means that each partition piece $\bar{\mathcal{F}}_{q,j}$ is the intersection between $\bar{\mathcal{F}}^*_{q_0,k_{q_0}}$, $\bar{\mathcal{F}}^*_{q_0+1,k_{q_0+1}}$, ..., $\bar{\mathcal{F}}^*_{q_0+1,k_q}$ for $k_{q_0} \in \left[1\colon N_{q_0}\right]$, $k_{q_0+1} \in \left[1\colon N_{q_0}+1\right]$, ..., $k_q \in \left[1\colon N_q\right]$. This means that there are $N_q \triangleq \prod_{p=q_0}^{q} N^*_p$ total partition pieces (for our constructed nested partitions) at this level.

Recall the potentially non-nested partitions were defined by bracketing functions $\left\{\left[l^*_{q,j}, u^*_{q,j}\right]\right\}_{j=1}^{N^*_q}$. Due to how we constructed the nested partitions $\left\{\left[l_{q,j}, u_{q,j}\right]\right\}_{j=1}^{N_q}$ (procedure described above), we have the following results:

- The nested partitions $\left\{\left[l_{q,j}, u_{q,j}\right]\right\}_{j=1}^{N_q}$ must be covering since the non-nested prtitions $\left\{\left[l^*_{q,j}, u^*_{q,j}\right]\right\}_{j=1}^{N^*_q}$ are covering and we took all possible intersections of these non-nested partitions of size $p \in [q_0 : q]$ to construct our nested partitions.

- The nested partitions $\left\{\left[l_{q,j}, u_{q,j}\right]\right\}_{j=1}^{N_q}$ must be at most of the size of the largest bracket out of $\left\{\left[l^*_{q,j}, u^*_{q,j}\right]\right\}_{j=1}^{N^*_q}$ again since we took all possible intersections of these non-nested partitions of size $p \in [q_0 : q]$ to construct our nested partitions. Thus, we define $\Delta_{q,j} \triangleq u^*_{q,j} - l^*_{q,j}$; note that this choice of $\Delta_{q,j}$ satisfies the conditions of display (C.8.5) since

$$\begin{aligned}\mathbb{E}_{\pi^*_{2:T}}\left[\left\{\pi^{*,(i)}_{2:T}\right\}^{-1}\Delta_{q,j}(\mathcal{H}^{(i)}_T)^2\right] &\leq \pi^{-(T-1)}_{\min}\mathbb{E}_{\pi^*_{2:T}}\left[\Delta_{q,j}(\mathcal{H}^{(i)}_T)^2\right] \\ &= \pi^{-(T-1)}_{\min}\mathbb{E}_{\pi^*_{2:T}}\left[\left(u^*_{q,j}(\mathcal{H}^{(i)}_T) - l^*_{q,j}(\mathcal{H}^{(i)}_T)\right)^2\right] \leq 2^{-2q}.\end{aligned}$$

  The first inequality above holds by Condition 5.3.2 and the second inequality holds by how we defined the non-nested partitions in display (C.8.18).

We now show that display (C.8.6) holds, i.e., the number of sets in the partition grows at a bounded rate as the size of the partition pieces goes to zero:

$$\sum_{q=q_0}^{\infty} 2^{-q}\sqrt{\log N_q} = \sum_{q=q_0}^{\infty} 2^{-q}\sqrt{\log\left(\prod_{p=q_0}^{q} N^*_p\right)} = \sum_{q=q_0}^{\infty} 2^{-q}\sqrt{\sum_{p=q_0}^{q}\log N^*_p}$$

Note that $\sqrt{\sum_{p=q_0}^{q}\log N^*_p} \leq \sum_{p=q_0}^{q}\sqrt{\log N^*_p}$ because $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any

positive non-negative values a, b.

$$\leq \sum_{q=q_0}^{\infty} 2^{-q} \sum_{p=q_0}^{q} \sqrt{\log N_p^*} = \sum_{q=q_0}^{\infty} 2^{-q} \sum_{p=q_0}^{\infty} \mathbb{I}_{p \leq q} \sqrt{\log N_p^*} = \sum_{p=q_0}^{\infty} \sqrt{\log N_p^*} \sum_{q=q_0}^{\infty} 2^{-q} \mathbb{I}_{p \leq q}$$

For the last equality above, we can exchange the infinite summations above by Fubini's theorem because the following argument will show that $\sum_{p=q_0}^{\infty} \sqrt{\log N_p^*} \sum_{q=q_0}^{\infty} 2^{-q} \mathbb{I}_{p \leq q}$ is bounded.

Since $\sum_{q=q_0}^{\infty} 2^{-q} \mathbb{I}_{p \leq q} = \sum_{q=p}^{\infty} 2^{-q} = 2^{-(p-1)}$,

$$= \sum_{p=q_0}^{\infty} 2^{-(p-1)} \sqrt{\log N_p^*} = 4 \sum_{p=q_0}^{\infty} 2^{-(p+1)} \sqrt{\log N_p^*}$$

$$= 4 \sum_{p=q_0}^{\infty} 2^{-(p+1)} \sqrt{\log N_{[\,]}\left(2^{-p} \pi_{\min}^{(T-1)/2}, \bar{\mathcal{F}}, L_2(\mathcal{P}_{\pi^*})\right)}$$

Since $N_{[\,]}\left(2^{-p} \pi_{\min}^{(T-1)/2}, \bar{\mathcal{F}}, L_2(\mathcal{P}_{\pi^*})\right)$ is monotonically increasing as p increases by lower Darboux sums, we have the following upper bound:

$$\leq 4 \int_0^{2^{-q_0}} \sqrt{\log N_{[\,]}\left(\varepsilon \pi_{\min}^{(T-1)/2}, \bar{\mathcal{F}}, L_2(\mathcal{P}_{\pi^*})\right)} d\varepsilon$$

Since we chose q_0 such that $\delta < 2^{-q_0} \leq 2\delta$,

$$
\begin{aligned}
&\leq 4 \int_0^{\delta} \sqrt{\log N_{[\,]}\big(\varepsilon \pi_{\min}^{(T-1)/2}, \bar{\mathcal{F}}, L_2(\mathcal{P}_{\pi^*})\big)}\, d\varepsilon \\
&\qquad \underbrace{\leq}_{(a)} 4 \int_0^{\delta} \sqrt{\log N_{[\,]}\big(\varepsilon \pi_{\min}^{(T-1)/2}, \mathcal{F}, L_2(\mathcal{P}_{\pi^*})\big)}\, d\varepsilon \\
&\qquad = 4 \pi_{\min}^{-(T-1)/2} \int_0^{\delta} \sqrt{\log N_{[\,]}\big(\varepsilon \pi_{\min}^{(T-1)/2}, \mathcal{F}, L_2(\mathcal{P}_{\pi^*})\big)}\, \pi_{\min}^{(T-1)/2} d\varepsilon \\
&\qquad\qquad \underbrace{=}_{(b)} 4 \pi_{\min}^{-(T-1)/2} \int_0^{\delta \pi_{\min}^{(T-1)/2}} \sqrt{\log N_{[\,]}\big(u, \mathcal{F}, L_2(\mathcal{P}_{\pi^*})\big)}\, du \underbrace{<}_{(c)} \infty.
\end{aligned}
$$

Inequality (a) above holds by display (C.8.3).

Equality (b) above holds by integration by substitution for $u = \pi_{\min}^{(T-1)/2} \varepsilon$.

Inequality (c) above holds by our finite bracketing integral assumption. ■

References

[1] (2017). *Sugars and dental caries*. Technical report, World Health Organization.

[2] Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems* (pp. 2312–2320).

[3] Agarwal, A., Agarwal, S., Assadi, S., & Khanna, S. (2017). Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory* (pp. 39–75).

[4] Agrawal, S. & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning* (pp. 127–135).

[5] Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

[6] Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

[7] Andrews, D. W. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric theory*, 4(3), 458–467.

[8] Andrews, I., Kitagawa, T., & McCloskey, A. (2019). *Inference on winners*. Technical report, National Bureau of Economic Research.

[9] Asadi, K. & Littman, M. L. (2017). An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning* (pp. 243–252).: PMLR.

[10] Bernstein, D. S. (2018). Scalar, vector, and matrix mathematics. In *Scalar, Vector, and Matrix Mathematics*. Princeton university press.

[11] Bibaut, A., Dimakopoulou, M., Kallus, N., Chambaz, A., & van der Laan, M. (2021a). Post-contextual-bandit inference. *Advances in Neural Information Processing Systems*, 34, 28548–28559.

[12] Bibaut, A., Kallus, N., Dimakopoulou, M., Chambaz, A., & van der Laan, M. (2021b). Risk minimization from adaptively collected data: Guarantees for supervised and policy learning. *Advances in Neural Information Processing Systems*, 34, 19261–19273.

[13] Boruvka, A., Almirall, D., Witkiewitz, K., & Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523), 1112–1121.

[14] Brannath, W., Gutjahr, G., & Bauer, P. (2012). Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association*, 107(498), 824–832.

[15] Brennan, J., Vinayak, R. K., & Jamieson, K. (2020). Estimating the number and effect sizes of non-null hypotheses. In *International Conference on Machine Learning* (pp. 1123–1133).: PMLR.

[16] Bubeck, S., Cesa-Bianchi, N., & Kakade, S. M. (2012). Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory* (pp. 41–1).: JMLR Workshop and Conference Proceedings.

[17] Bubeck, S., Munos, R., & Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory* (pp. 23–37).: Springer.

[18] Bura, E., Duarte, S., Forzani, L., Smucler, E., & Sued, M. (2018). Asymptotic theory for maximum likelihood estimates in reduced-rank multivariate generalized linear models. *Statistics*, 52(5), 1005–1024.

[19] Caria, S., Kasy, M., Quinn, S., Shami, S., & Teytelboym, A. (2020). An adaptive targeted field experiment: Job search assistance for refugees in jordan. *CESifo Working Paper*.

[20] Cesa-Bianchi, N., Gentile, C., Lugosi, G., & Neu, G. (2017). Boltzmann exploration done right. *Advances in neural information processing systems*, 30.

[21] Cesa-Bianchi, N. & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

[22] Chandak, Y., Theocharous, G., Shankar, S., White, M., Mahadevan, S., & Thomas, P. (2020). Optimizing for the future in non-stationary mdps. In *International Conference on Machine Learning* (pp. 1414–1425).: PMLR.

[23] Chen, H., Lu, W., & Song, R. (2020). Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, (pp. 1–16).

[24] Dean, S., Mania, H., Matni, N., Recht, B., & Tu, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*.

[25] Deshpande, Y., Javanmard, A., & Mehrabi, M. (2019). Online debiasing for adaptively collected high-dimensional data. *arXiv preprint arXiv:1911.01040*.

[26] Deshpande, Y., Mackey, L., Syrgkanis, V., & Taddy, M. (2018). Accurate inference for adaptive linear models. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research* (pp. 1194–1203). Stockholmsmässan, Stockholm Sweden: PMLR.

[27] Druce, K. L., Dixon, W. G., & McBeth, J. (2019). Maximizing engagement in mobile health studies: lessons learned and future directions. *Rheumatic Disease Clinics*, 45(2), 159–172.

[28] Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

[29] Dvoretzky, A. (1972). Asymptotic normality for sums of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*: The Regents of the University of California.

[30] Dye, B. A., Thornton-Evans, G., Li, X., & Iafolla, T. (2015). *Dental caries and tooth loss in adults in the United States, 2011-2012*. US Department of Health and Human Services, Centers for Disease Control and Prevention.

[31] Engle, R. & McFadden, D. (1994). *Handbook of Econometrics*. Technical report, Elsevier.

[32] Erraqabi, A., Lazaric, A., Valko, M., Brunskill, E., & Liu, Y.-E. (2017). Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics* (pp. 709–717).: PMLR.

[33] Eysenbach, G. (2005). The law of attrition. *Journal of medical Internet research*, 7(1), e11.

[34] Figueroa, C. A., Aguilera, A., Chakraborty, B., Modiri, A., Aggarwal, J., Deliu, N., Sarkar, U., Jay Williams, J., & Lyles, C. R. (2021). Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. *Journal of the American Medical Informatics Association*, 28(6), 1225–1234.

[35] Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

[36] Forman, E. M., Kerrigan, S. G., Butryn, M. L., Juarascio, A. S., Manasse, S. M., Ontañón, S., Dallal, D. H., Crochiere, R. J., & Moskow, D. (2019). Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of behavioral medicine*, 42(2), 276–290.

[37] Gao, Z., Han, Y., Ren, Z., & Zhou, Z. (2019). Batched multi-armed bandits problem. *Conference on Neural Information Processing Systems*.

[38] Ghosh, S., Kim, R., Chhabria, P., Dwivedi, R., Klasjna, P., Liao, P., Zhang, K., & Murphy, S. (2023). Did we personalize? assessing personalization by an online reinforcement learning algorithm using resampling. *arXiv preprint arXiv:2304.05365*.

[39] Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., & Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15).

[40] Hammersley, J. (2013). *Monte carlo methods*. Springer Science & Business Media.

[41] Han, Y., Zhou, Z., Zhou, Z., Blanchet, J., Glynn, P. W., & Ye, Y. (2020). Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*.

[42] Hazelton, M. L. (2011). *Methods of Moments Estimation*, (pp. 816–817). Springer Berlin Heidelberg: Berlin, Heidelberg.

[43] Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 1055 – 1080.

[44] Hu, F. & Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*, volume 525. John Wiley & Sons.

[45] Huber, P. J. (1967). Under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*, volume 5 (pp. 221).: Univ of California Press.

[46] Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics* (pp. 492–518). Springer.

[47] Hung, K. & Fithian, W. (2019). Rank verification for exponential families. *Annals of Statistics*.

[48] Imbens, G. W. & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[49] Jamieson, K., Malloy, M., Nowak, R., & Bubeck, S. (2014). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory* (pp. 423–439).

[50] Jamieson, K. & Nowak, R. (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)* (pp. 1–6).: IEEE.

[51] Jiang, N. & Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning* (pp. 652–661).: PMLR.

[52] Jun, K.-S., Jamieson, K. G., Nowak, R. D., & Zhu, X. (2016). Top arm identification in multi-armed bandits with batch arm pulls. In *AISTATS* (pp. 139–148).

[53] Kallus, N. & Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167).

[54] Karampatziakis, N., Mineiro, P., & Ramdas, A. (2021). Off-policy confidence sequences. In *International Conference on Machine Learning* (pp. 5301–5310).: PMLR.

[55] Kasy, M. (2019). Uniformity and the delta method. *Journal of Econometric Methods*, 8(1).

[56] Kasy, M. & Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1), 113–132.

[57] Kaufmann, E. & Koolen, W. (2018). Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv preprint arXiv:1811.11419*.

[58] Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170–179).

[59] Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, 117(26), 14900–14905.

[60] Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., & Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S), 1220.

[61] Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., & Murphy, S. A. (2019). Efficacy of contextually tailored suggestions for physical activity: A microrandomized optimization trial of heartsteps. *Annals of Behavioral Medicine*, 53(6), 573–582.

[62] Lai, T. L. & Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1), 154–166.

[63] Lattimore, T. & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

[64] Leeb, H. & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, (pp. 21–59).

[65] Li, C. & De Rijke, M. (2019). Cascading non-stationary bandits: Online learning to rank in the non-stationary cascade model. *arXiv preprint arXiv:1905.12370*.

[66] Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web* (pp. 661–670).

[67] Liao, P., Greenewald, K., Klasnja, P., & Murphy, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1), 1–22.

[68] Liao, P., Klasnja, P., Tewari, A., & Murphy, S. A. (2016). Sample size calculations for micro-randomized trials in mhealth. *Statistics in medicine*, 35(12), 1944–1971.

[69] Liu, Q., Proschan, M. A., & Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association*, 97(460), 1034–1041.

[70] Liu, Y.-E., Mandel, T., Brunskill, E., & Popovic, Z. (2014). Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM* (pp. 161–168).

[71] Luedtke, A. R. & Laan, M. J. v. d. (2018). Parametric-rate inference for one-sided differentiable parameters. *Journal of the American Statistical Association*, 113(522), 780–788.

[72] Luedtke, A. R. & Van Der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2), 713.

[73] Mate, A., Wilder, B., Taneja, A., & Tambe, M. (2023). Improved policy evaluation for randomized trials of algorithmic resource allocation. *International Conference on Machine Learning*.

[74] Nahum-Shani, I., Greer, Z. M., Trella, A. L., **Zhang**, t., Carpenter, S., Elashoff, D., Murphy, S. A., & Shetty, V. (2023). Optimizing an adaptive digital oral health intervention for promoting oral self care behaviors: Micro-randomized trial protocol. *Preparing for submission*.

[75] Nickerson, D. M. (1994). Construction of a conservative confidence region from projections of an exact confidence region in multiple linear regression. *The American Statistician*, 48(2), 120–124.

[76] Nie, X., Tian, X., Taylor, J., & Zou, J. (2018). Why adaptively collected data have negative bias and how to correct for it. *International Conference on Artificial Intelligence and Statistics*.

[77] Offer-Westort, M., Coppock, A., & Green, D. P. (2019). Adaptive experimental design: Prospects and applications in political science. *Available at SSRN 3364402*.

[78] Perchet, V., Rigollet, P., Chassang, S., Snowberg, E., et al. (2016). Batched bandit problems. *The Annals of Statistics*, 44(2), 660–681.

[79] Piette, J. D., Newman, S., Krein, S. L., Marinec, N., Chen, J., Williams, D. A., Edmond, S. N., Driscoll, M., LaChappelle, K. M., Maly, M., et al. (2022). Artificial intelligence (ai) to improve chronic pain care: Evidence of ai learning. *Intelligence-Based Medicine*, (pp. 100064).

[80] Qian, T., Walton, A. E., Collins, L. M., Klasnja, P., Lanza, S. T., Nahum-Shani, I., Rabbi, M., Russell, M. A., Walton, M. A., Yoo, H., et al. (2022). The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations. *Psychological methods*.

[81] Qian, T., Yoo, H., Klasnja, P., Almirall, D., & Murphy, S. A. (2021). Estimating time-varying causal excursion effects in mobile health with binary outcomes. *Biometrika*, 108(3), 507–527.

[82] Rafferty, A., Ying, H., & Williams, J. (2019). Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *JEDM| Journal of Educational Data Mining*, 11(1), 47–79.

[83] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12), 1393–1512.

[84] Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality* (pp. 69–117). Springer.

[85] Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

[86] Romano, J. P., Shaikh, A. M., et al. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, 40(6), 2798–2822.

[87] Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4), 500–522.

[88] Shaikh, H., Modiri, A., Williams, J. J., & Rafferty, A. N. (2019). Balancing student success and inferring personalized effects in dynamic experiments. In *EDM*.

[89] Shin, J., Ramdas, A., & Rinaldo, A. (2019). Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems* (pp. 7100–7109).

[90] Simchi-Levi, D. & Wang, C. (2023). Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *International Conference on Artificial Intelligence and Statistics* (pp. 3086–3097).: PMLR.

[91] Stanford Computational Policy Lab (2023). Encouraging appearance in court using new techniques from reinforcement learning.

[92] Sutton, R. S. & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[93] Tang, L., Jiang, Y., Li, L., & Li, T. (2014). Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems* (pp. 73–80).

[94] Thomas, P. & Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning* (pp. 2139–2148).: PMLR.

[95] Tomkins, S., Liao, P., Klasnja, P., & Murphy, S. (2021). Intelligentpooling: Practical thompson sampling for mhealth. *Machine learning*, 110(9), 2685–2727.

[96] Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., & Murphy, S. A. (2022). Designing reinforcement learning algorithms for digital interventions: Pre-implementation guidelines. *Algorithms*, 15(8).

[97] Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., & Murphy, S. A. (2023). Reward design for an online reinforcement learning algorithm supporting oral self-care. *Thirty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-23)*.

[98] Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

[99] Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

[100] Van Der Vaart, A. W. & Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes* (pp. 16–28). Springer.

[101] Villar, S. S., Bowden, J., & Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2), 199.

[102] Wang, Y.-X., Agarwal, A., & Dudık, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning* (pp. 3589–3597).: PMLR.

[103] Wassmer, G. & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Springer.

[104] Yao, J., Brunskill, E., Pan, W., Murphy, S., & Doshi-Velez, F. (2021). Power constrained bandits. In K. Jung, S. Yeung, M. Sendak, M. Sjoding, & R. Ranganath (Eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research* (pp. 209–259).: PMLR.

[105] Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M., & Hochberg, I. (2017). Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10), e338.

[106] Zeger, S. L. & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, (pp. 121–130).

[107] Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(1), 1–16.

[108] Zhan, R., Hadad, V., Hirshberg, D. A., & Athey, S. (2021). Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 2125–2135).

[109] Zhang, K. W., Janson, L., & Murphy, S. A. (2020). Inference for batched bandits. *Advances in neural information processing systems*, 33, 9818–9829.

[110] Zhang, K. W., Janson, L., & Murphy, S. A. (2021). Statistical inference with m-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34, 7460–7471.

[111] Zhang, K. W., Janson, L., & Murphy, S. A. (2022). Statistical inference after adaptive sampling for longitudinal data. *arXiv preprint arXiv:2202.07098*.