Estimating Time-Varying Causal Excursion Effect in Mobile Health with Binary Outcomes

Tianchen Qian*, Hyesun Yoo†, Predrag Klasnja‡, Daniel Almirall§, and Susan A. Murphy¶

June 4, 2019

Abstract

Advances in wearables and digital technology now make it possible to deliver behavioral mobile health interventions to individuals in their everyday life. The micro-randomized trial (MRT) is increasingly used to provide data to inform the construction of these interventions. This work is motivated by multiple MRTs that have been conducted or are currently in the field in which the primary outcome is a longitudinal binary outcome. The first, often called the primary, analysis in these trials is a marginal analysis that seeks to answer whether the data indicates that a particular intervention component has an effect on the longitudinal binary outcome. Under rather restrictive assumptions one can, based on existing literature, derive a semiparametric, locally efficient estimator of the causal effect. In this paper, starting from this estimator, we develop an estimator that can be used as the basis of a primary analysis under more plausible assumptions. Simulation studies are conducted to compare the estimators. We illustrate the developed methods using data from the MRT, BariFit. In BariFit, the goal is to support weight maintenance for individuals who received bariatric surgery.

Keywords: Micro-randomized trials; Binary outcome; Longitudinal data; Causal inference; Causal excursion effect; Semiparametric efficiency; Mobile health.
1 Introduction

In mobile health (mHealth), mobile devices (including smart phones and wearable devices) are used to deliver interventions intended to promote healthy behaviors and health-related behavioral change (Free et al. 2013). Treatments include prompts to self-monitor, cognitive interventions to promote reflection and goal setting as well as suggestions of ways to enact healthy behavior changes. These treatments are delivered to the individual via the individual’s phone or a wearable. An increasingly common trial, called the micro-randomized trial (MRT), is being used to inform the development of mHealth interventions (Liao et al. 2016, Klasnja et al. 2015, Law et al. 2016, Klasnja et al. 2018, Kramer et al. 2019, Tate & Valle 2019). In an MRT, each individual is repeatedly randomized among the multiple options for a treatment, often hundreds or even thousands of times over the course of the trial. In all cases the randomization probabilities are determined as part of the design of the trial and are thus known. Between randomizations, covariate data is collected on the individual’s current/recent context via sensors and/or self-report, and after each randomization a “proximal,” near-time outcome is collected. The time-varying treatments and covariates as well as the proximal outcome comprise the longitudinal data for use in assessing if a treatment has an effect on the proximal outcome and/or in which settings this effect may be greater or smaller. Such knowledge is crucial for informing decisions regarding whether to include the treatment in the mHealth intervention as well as garnering an understanding of the contexts in which the treatment might be more effective.

This paper is motivated by our involvement in a number of MRTs in which the primary proximal outcome is binary. Schematics of these trials can be found at the website of the Methodology Center at the Pennsylvania State University1. For example, in the Substance Abuse Research Assistance study (Rabbi et al. 2018), one goal is to learn whether sending the user a “present” such as an inspirational quote by a popular celebrity will increase the user’s evening self-report completion rate. In another MRT, currently in the field, Smart

1https://methodology.psu.edu/ra/adap-inter/mrt-projects/#proj
Weight Loss Management as well as in the MRT conducted by JOOL Health (Bidargaddi et al. 2018), and in the BariFit (Ridpath 2017) MRT, one goal is to determine whether tailored reminder messages will differentially influence daily self-monitoring. In all of these cases the primary proximal outcome is the binary outcome of whether or not self-monitoring is completed, and there is a need for data analysis methods for use in conducting both primary as well as secondary data analyses for time-varying binary proximal outcomes in MRTs.

Most mHealth treatments that might be delivered many times (e.g. daily or within a day) have been designed to have their greatest impact on a near time, proximal outcome (Heron & Smyth 2010). As a result, primary analyses for these treatments focus on effects on the proximal outcome. Secondary analyses for MRTs usually concern treatment effect modification; i.e., the interest is in assessing whether the magnitude of the treatment effect depends on certain variables (such as time or certain contextual information that can be time-varying).

In this paper, we consider inference for causal effects that can be used as the basis of these primary and secondary analyses. One possible causal effect is akin to Robins’ treatment “blip” in the structural nested mean model for binary outcomes (Robins 1994, 1997). Our first contribution is that we consider marginal generalizations of this effect which we call causal excursion effects. Such effects can be viewed as “excursions” as they represent a causal effect of a treatment occurring over an interval of time extending into the future. In this case the definition of the excursion effect involves rules for how further treatments, if any, would occur during this interval of time. This is well suited for answering questions that naturally arise in MRTs such as “what is the effect of delivering a treatment now then not delivering any treatment for the next \( m \) time points”. Furthermore, these causal effects may be moderated by past treatment, thus these effects might be interpreted as contrasts between excursions from the treatment protocol as specified by the micro-randomization. Lastly, causal excursion effects are often marginal in that the effect is defined as marginal over all but a small subset of the individual’s prior data. See Section 2

\(^2\)http://grantome.com/grant/NIH/R01-DK108678-01A1
We, based on Robins (1994), provide a semiparametric, locally efficient estimator of one possible causal excursion effect. In this case the model for the causal effect is conditional on the entire history, i.e., all variables that occur prior to the current time. Our second contribution is that we, starting from this estimator, develop an estimator that consistently estimates the causal effect conditional on an arbitrary subset of the history. The estimator is robust in the sense that, for consistency, it does not require that the model for the proximal outcome under no treatment to be correctly specified. We propose to use this estimator as the basis of primary analyses and secondary moderation analyses of MRTs with binary proximal outcomes.

2 Preliminaries

2.1 Micro-randomized trials and BariFit

As introduced in Section 1, micro-randomized trials (MRTs) provide longitudinal data for use in developing mHealth interventions (Liao et al. 2016, Dempsey et al. 2015, Klasnja et al. 2015). BariFit, for example, is an MRT that was conducted to aid in the process of developing an mHealth intervention for promoting weight maintenance among individuals who received bariatric surgery (Ridpath 2017). In this study a daily text reminder might be sent to encourage the participant to self-monitor his/her food intake via a food log; we will refer to this daily text reminder as the food track reminder.

In an MRT each participant is randomized, with known probabilities, between the treatment options at predetermined time points. In BariFit, the food track reminder is randomized with probability 0.5 between deliver versus do not deliver every morning for 112 days. In general, the randomization probability can vary depending on the individual’s data observed up to that time.

In BariFit, the proximal outcome for the food track reminder is whether the participant completes his/her food log on that day. The analysis method developed here focuses on
this proximal outcome. However, hopefully these reminders will assist the individual in building up healthy habits, so longer term effects are desired. Thus, in defining the causal effects below, we do not assume that longer term effects are absent.

Because treatments are delivered to individuals during their everyday life, there may be unethical or unsafe times at which it is inappropriate or deemed excessively burdensome to deliver a treatment. For example, if the treatment is a smartphone notification that audibly pings and makes the phone light up, it is inappropriate to deliver the smartphone notification when the individual might be operating a motor vehicle (Klasnja et al. 2018). In these cases, randomization occurs only at available time points, and the causal effect is conditional on the available times (Boruvka et al. 2018). Due to the fact that many MRTs involve considerations of availability, the methods developed below accommodate this. However, in the case of the BariFit food track reminders, they were sent, if at all, early in the morning and, as text messages remain on the phone, the participant is able to read them at a time s/he deems convenient. Thus, in the BariFit study, lack of availability is not a consideration.

2.2 Related literature and our contribution

As data from mHealth studies are often longitudinal, generalized estimating equations (Liang & Zeger 1986) and random effects models (Laird & Ware 1982) are the most commonly used methods for modeling the time-varying association between two or more variables in mHealth studies (Schwartz & Stone 2007, Bolger & Laurenceau 2013). However, in the presence of time-varying treatment or time-varying covariates, it is well known that the use of these methods can result in biased causal effect estimates without strong and often unrealistic assumptions (Pepe & Anderson 1994, Schildcrout & Heagerty 2005). Structural nested mean models (SNMMs) and marginal structural models (MSMs) are two classes of models that facilitate estimation of causal effects of a time-varying treatment on a time-varying outcome, where the treatment assignment mechanism may depend on history covariates (Robins 1994, 1997, 2000, Robins et al. 2000). In a SNMM, the effect of
sequentially removing an amount of treatment on future outcomes, after having removed all future treatments, is modeled. This effect is a conditional expectation given all the history information up to that time. In a MSM, the expectation of the time-varying outcome under a fixed treatment trajectory (possibly conditional on a subset of baseline covariates) is modeled as a function of the treatment trajectory and the subset of baseline covariates.

The causal excursion effect we considered can be conditional on an arbitrary subset of the history. Unlike MSM, our approach allows estimation of causal effect modification by time-varying covariates. Unlike SNMM, our causal excursion effect is marginalized over variables not in the subset of interested variables, i.e., possibly marginal over a large part of the treatment and covariate histories. This makes the estimand coherent with the goal of primary and secondary analyses, and avoids modeling the relationship between the time-varying outcome and the past history in MRTs, where the number of time points can be numerous and the history can be high-dimensional. A related marginalization idea was considered by Neugebauer et al. (2007) in the “history-restricted” extension of MSM. Furthermore, the causal excursion effect can be defined as a contrast between two treatment excursions extended into the future; this excursion aspect was not considered in either SNMM or MSM.

In the previous work on data analytic methods for MRTs, Boruvka et al. (2018) and Dempsey et al. (2017) considered estimation of causal effects of mHealth interventions, where the outcome is continuous. In this paper we consider binary outcome, and we address the unique challenges raised by the binary outcome by considering a log relative risk model for the causal excursion effect and by developing a novel estimator.

3 Definition and assumptions

3.1 Notation and observed data

Suppose that for each individual, there are $T$ time points at which the treatment can be delivered ($T$ need not be the same for each individual). For simplicity we assume that
there are two treatment options which we will call treatment and no treatment. Thus, the treatment assignment at time \( t \), \( A_t \), is binary, where 1 means treatment and 0 means no treatment. Denote by \( X_t \) the vector of observations collected after time \( t-1 \) and up to/including time \( t \); \( X_1 \) includes baseline covariates. \( X_t \) contains the availability indicator, \( I_t \): \( I_t = 1 \) if the individual is available for treatment at time \( t \) and \( I_t = 0 \) otherwise. If \( I_t = 0 \), randomization will not occur at time \( t \) and \( A_t = 0 \). We use overbar to denote a sequence of variables up to a decision point; for example \( \bar{A}_t = (A_1, \ldots, A_t) \). Information accrued up to time \( t \) is represented by the history \( H_t = (X_1, A_1, X_2, A_2, \ldots, X_{t-1}, A_{t-1}, X_t) = (\bar{X}_t, \bar{A}_{t-1}) \). The randomization probability for \( A_t \) can depend on \( H_t \), and is denoted by \( p_t(H_t) = P(A_t = 1 | H_t) \); \( p_t(\cdot) \) is known by the MRT design. The observed data on a generic individual, ordered in time, is \( O = (X_1, A_1, \ldots, X_T, A_T, X_{T+1}) \). We assume that the data from different individuals are independent and identically distributed draws from an unknown distribution \( P_0 \). Unless noted otherwise, all expectations are taken with respect to \( P_0 \).

The proximal outcome, \( Y_{t,\Delta} \), following the treatment assignment at time \( t \), is a known function of the individual’s data within a subsequent window of length \( \Delta \), where \( \Delta \geq 1 \) is a positive integer; i.e., \( Y_{t,\Delta} = y(X_{t+1}, A_{t+1}, \ldots, X_{t+\Delta-1}, A_{t+\Delta-1}, X_{t+\Delta}) \) for some known function \( y(\cdot) \). In this paper \( Y_{t,\Delta} \) is binary. For example, in a smoking cessation study where the treatment is a push notification that reminds the user to practice stress-reduction exercises [Spring 2017], the treatment is randomized every minute (albeit with very low probability of sending a push notification at any given minute), and the proximal outcome is whether the user experiences a stress episode during the 120-minute window following a treatment. In this example, \( t \) is every minute, and \( \Delta = 120 \). A simpler setting with \( \Delta = 1 \) is where the proximal outcome cannot depend on future treatment and is given by \( Y_{t,1} = y(X_{t+1}) \); an example is the BariFit MRT described in Section 2.1, where the randomization once a day, and the proximal outcome is measured within each day. The estimator we propose in Section 5 allows for general \( \Delta \).

For an arbitrary function \( f(\cdot) \) of the generic observed data \( O \), denote by \( \mathbb{P}_n f(O) \) the sample average \( \frac{1}{n} \sum_{i=1}^{n} f(O_i) \) where \( O_i \) denotes the \( i \)th individual’s observed data. We omit the subscript \( i \) for the \( i \)th individual throughout the paper unless necessary. We use \( \mathbb{1}(\cdot) \)
3.2 Potential outcomes and causal excursion effect

To define treatment effects, we use the potential outcomes framework (Rubin 1974, Robins 1986). For an individual, let $X_t(\bar{a}_{t-1})$ and $A_t(\bar{a}_{t-1})$ be the observation that would have been observed and the $t$th treatment that would have been assigned, respectively, if s/he were assigned the treatment sequence $\bar{a}_{t-1}$. Then the potential outcomes are defined as

$$\{X_1, A_1, X_2(a_1), A_2(a_1), X_3(\bar{a}_2), \ldots, X_{T+1}(\bar{a}_T) \text{ for all } \bar{a}_T \in \{0,1\}^{\otimes T}\}, \quad (1)$$

where $\otimes$ denotes the Cartesian product. The potential outcome for the proximal outcome is $Y_t, \Delta(\bar{a}_{t+\Delta-1})$. The treatment at time $t$ in (1) is indexed by past treatments because in an MRT the randomization probabilities can depend on the participant’s past treatment. However for notational simplicity, which will be further justified by Assumption 1 in Section 3.3, henceforth denote $A_2(A_1)$ by $A_2$ and so on with $A_t(\bar{A}_{t-1})$ by $A_t$.

The potential history under the observed treatment sequence at time $t$ is $H_t(\bar{A}_{t-1}) = (X_1, A_1, X_2(A_1), A_2, X_3(\bar{A}_2), \ldots, X_t(\bar{A}_{t-1}))$.

We define the causal effect of $A_t$ on $Y_t, \Delta$ using the log relative risk scale:

$$\beta_M\{t,S_t(\bar{A}_{t-1})\} = \log \frac{E\{Y_{t,\Delta}(\bar{A}_{t-1},1,\bar{0}) | S_t(\bar{A}_{t-1}), I_t(\bar{A}_{t-1}) = 1\}}{E\{Y_{t,\Delta}(\bar{A}_{t-1},0,\bar{0}) | S_t(\bar{A}_{t-1}), I_t(\bar{A}_{t-1}) = 1\}}, \quad (2)$$

where $S_t(\bar{A}_{t-1})$ is a vector of summary variables formed from $H_t(\bar{A}_{t-1})$, and $\bar{0}$ is a vector of length $\Delta - 1$. Expression (2) denotes the contrast of the expected outcome under two “excursions”: treatment at time $t$ and no treatment for the next $\Delta - 1$ time points, versus no treatment at time $t$ and no treatment for the next $\Delta - 1$ time points. We call $\beta_M\{t,S_t(\bar{A}_{t-1})\}$ a causal excursion effect. The expectation in (2) marginalizes over the randomization distribution of $\bar{A}_{t-1}$ that are not included in $S_t(\bar{A}_{t-1})$. In other words, the meaning of the causal excursion is relative to how treatment was assigned in the past: at time $t$, we are considering excursions from the current protocol of assigning treatment. The methods developed below generalize to other types of excursions, such as excursions that specify a decision rule at each time between time $t$ and time $t + \Delta - 1$. 

8
When $\Delta = 1$, (2) is a marginal generalization of the treatment “blip” in structural nested mean models (Robins 1994, 1997); hence the subscript “M” in $\beta_M\{t,S_t(\bar{A}_{t-1})\}$. A special case of $\beta_M\{t,S_t(\bar{A}_{t-1})\}$ is one of the blips considered by Robins (1994) given by

$$\beta_C\{t,H_t(\bar{A}_{t-1})\} = \log \frac{E\{Y_{t,1}(\bar{A}_{t-1},1) \mid H_t(\bar{A}_{t-1}),I_t(\bar{A}_{t-1}) = 1\}}{E\{Y_{t,1}(\bar{A}_{t-1},0) \mid H_t(\bar{A}_{t-1}),I_t(\bar{A}_{t-1}) = 1\}}.$$  (3)

In (2) we allow $S_t(\bar{A}_{t-1})$ to be any proper subset of $H_t(\bar{A}_{t-1})$, because the primary, pre-specified analysis in an MRT usually aims to assess whether a particular intervention component has a marginal effect on the proximal outcome. For such analysis, one sets $S_t(\bar{A}_{t-1}) = \emptyset$; i.e., the treatment effect is fully marginal. Subsequent analyses usually have a hierarchy of increasingly complex $S_t(\bar{A}_{t-1})$, e.g., variables that may modify the treatment effect. In this paper we sometimes call $\beta_M\{t,S_t(\bar{A}_{t-1})\}$ a “marginal excursion effect” to emphasize its marginal aspect. The concept of marginalization over all but a subset of $H_t$ also appeared in the MSM literature (Neugebauer et al. 2007). See Section 8 for further discussion regarding the excursion aspect of the causal effect.

There has been much discussion over the choice of association measure for binary outcomes in the literature, and reasons to prefer relative risk (over odds ratio) include its interpretability and collapsibility (Greenland 1987, Lumley et al. 2006). A drawback of using the relative risk as opposed to odds ratio is that the relative risk does not ensure that the estimated probability of success lies in the interval $[0,1]$. Nonetheless, we chose to define (2) on the relative risk scale, both for interpretability and modeling ease. See Section 8 for further discussion concerning this modeling choice.

### 3.3 Identification of parameters

To express the causal excursion effect in terms of the observed data, we make the following assumptions.

---

In the following, because we will introduce notation such as $S_T^T \beta$, we will use $S_t = \emptyset$ (to emphasize that the treatment effect is fully marginal) and $S_t = 1$ (to emphasize that $S_T^T \beta$ only has an intercept term) interchangeably as long as no confusion is caused.
**Assumption 1** (Consistency). The observed data equals the potential outcome under observed treatment assignment. In particular, \( X_2 = X_2(A_1), A_2 = A_2(A_1) \), and for each subsequent \( t \leq T \), \( X_t = X_t(\bar{A}_{t-1}), A_t = A_t(\bar{A}_{t-1}) \), and lastly, \( X_{T+1} = X_{T+1}(\bar{A}_T) \). This implies \( Y_{t,\Delta} = Y_{t,\Delta}(\bar{A}_{t+\Delta-1}) \).

**Assumption 2** (Positivity). If \( \Pr(H_t = h_t, I_t = 1) > 0 \), then \( \Pr(A_t = a \mid H_t = h_t, I_t = 1) > 0 \) for \( a \in \{0, 1\} \).

**Assumption 3** (Sequential ignorability). For \( 1 \leq t \leq T \), the potential outcomes \( \{X_{t+1}(\bar{a}_t), A_{t+1}(\bar{a}_t), \ldots, X_{T+1}(\bar{a}_T) : \bar{a}_T \in \{0, 1\}^{\otimes T}\} \) are independent of \( A_t \) conditional on \( H_t \).

In an MRT, because the treatment is sequentially randomized with known probabilities bounded away from 0 and 1, Assumptions 2 and 3 are satisfied by design. Assumption 1 may fail to hold if there is peer influence or social interaction between individuals; for example, in mHealth interventions with social media components, one individual’s proximal outcome may be dependent on another individual’s treatment assignment, which violates Assumption 1. In those cases, a causal inference framework that incorporates interference needs to be used (Hong & Raudenbush 2006, Hudgens & Halloran 2008). To maintain the focus of this paper we do not consider such settings here.

We show in Appendix A that under Assumptions 1 - 3, the causal excursion effect (2) can be written in terms of the observed data distribution:

\[
\beta_M \{t, S_t(\bar{A}_{t-1})\} = \log \frac{E \left[ E \left\{ \prod_{j=t+1}^{t+\Delta-1} \frac{1_{(A_j=0)}}{1-p_j(H_j)} Y_{t,\Delta} \mid A_t = 1, H_t, I_t = 1 \right\} \bigg| S_t, I_t = 1 \right]}{E \left[ E \left\{ \prod_{j=t+1}^{t+\Delta-1} \frac{1_{(A_j=0)}}{1-p_j(H_j)} Y_{t,\Delta} \mid A_t = 0, H_t, I_t = 1 \right\} \bigg| S_t, I_t = 1 \right]},
\]

where we define \( \prod_{j=t+1}^{t+\Delta-1} \frac{1_{(A_j=0)}}{1-p_j(H_j)} = 1 \) if \( \Delta = 1 \). With a slight abuse of notation, we denote the right hand side of (4) by \( \beta_M(t, S_t) \). Similarly the treatment blip given in (3) can be written as

\[
\beta_C(t, H_t) = \log \frac{E(Y_{t,1} \mid A_t = 1, H_t, I_t = 1)}{E(Y_{t,1} \mid A_t = 0, H_t, I_t = 1)}.
\]
4 A semiparametric, locally efficient estimator

To motivate the estimator for the marginal excursion effect $\beta_M(t, S_t)$, we first consider the special case where the treatment effect is conditional on the entire history $H_t$ and the proximal outcome is defined with $\Delta = 1$; that is, consider (5). Using techniques in Robins (1994), the semiparametric efficient score (Newey 1990) can be derived; a proof is provided in Appendix F.

**Theorem 1.** Suppose $f(\cdot)$ is a known deterministic function such that for $1 \leq t \leq T$,

$$\beta_C(t, H_t) = f(H_t)^T \psi,$$

for some unknown value of a $p$-dimensional parameter $\psi$. In the semiparametric model characterized by (6) and Assumptions 1, 2 and 3, the efficient score for $\psi$ is

$$S_{\text{eff}}(\psi) = \sum_{t=1}^{T} I_t e^{-A_t f(H_t)^T \psi} \{Y_{t+1} - e^{\mu(H_t) + A_t f(H_t)^T \psi} \} K_t \{A_t - p_t(H_t)\} f(H_t),$$

where

$$\mu(H_t) = \log E(Y_{t+1} \mid H_t, A_t = 0),$$

$$K_t = \frac{e^{f(H_t)^T \psi} \{1 - e^{\mu(H_t)}\} p_t(H_t) + \{1 - e^{\mu(H_t)} + f(H_t)^T \psi\} \{1 - p_t(H_t)\}}{1 - e^{\mu(H_t)}}.$$

It follows from semiparametric efficiency theory that the solution $\hat{\psi}$ to $P_n S_{\text{eff}}(\psi) = 0$ achieves the semiparametric efficiency bound; i.e., it has the smallest asymptotic variance among all semiparametric regular and asymptotically linear estimators for $\psi$ (Newey 1990, Tsiatis 2007). Of course this estimator is not practical because $S_{\text{eff}}$ depends on an unknown quantity $\mu(H_t)$. In practice, one can replace $\mu(H_t)$ with a parametric working model and solve for the estimating equation. Because $S_{\text{eff}}(\psi)$ is robust to misspecified $\mu(H_t)$ (i.e., it has expectation 0 even if one replaces $\mu(H_t)$ by an arbitrary function of $H_t$), the resulting estimator is semiparametric locally efficient, in the sense that it is consistent and when the working model for $\mu(H_T)$ is correctly specified it attains the semiparametric efficiency bound.
Here we describe a particular implementation of this efficient score; this implementation serves to motivate the proposed method in Section 5, where we consider estimation of a causal excursion effect in which $\Delta \geq 1$ and for which the causal excursion effect is marginal. Let the working model for $\mu(H_t)$ be $g(H_t)^T \alpha$, where $g(H_t)$ is a vector of features constructed from $H_t$ and $\alpha$ is a finite dimensional parameter. We combine the resulting estimating function from (7) with an estimating function for $\alpha$ from $H_t$ and $\alpha$ is correct model for $g(H_t)$.

\begin{equation}
    m_C(\alpha, \psi) = \sum_{t=1}^{T} I_t e^{-A_t f(H_t)^T \psi} \{ Y_{t+1} - e^{g(H_t)^T \alpha + A_t f(H_t)^T \psi} \} \tilde{K}_t \left[ \frac{g(H_t)}{(A_t - p_t(H_t)) f(H_t)} \right],
\end{equation}

where

\[ \tilde{K}_t = \frac{e^{f(H_t)^T \psi}}{e^{f(H_t)^T \psi} \{ 1 - e^{g(H_t)^T \alpha} \} p_t(H_t) + \{ 1 - e^{g(H_t)^T \alpha + f(H_t)^T \psi} \} \{ 1 - p_t(H_t) \}}. \]

In Appendix B we prove the following result.

**Theorem 2.** Suppose (6) and Assumptions 1, 2 and 3 hold. Let $\dot{m}_C$ be the derivative of $m_C(\alpha, \psi)$ with respect to $(\alpha, \psi)$. Let $(\hat{\alpha}, \hat{\psi})$ be a solution to $P_n m_C(\alpha, \psi) = 0$. Suppose $\psi^*$ is the value of $\psi$ corresponding to the data generating distribution, $P_0$. Under regularity conditions, $\sqrt{n}(\hat{\psi} - \psi^*)$ is asymptotically normal with mean zero and variance-covariance matrix $\Sigma_C$. A consistent estimator for $\Sigma_C$ is the lower block diagonal $(p \times p)$ entry of the matrix $\{ P_n \dot{m}_C(\hat{\alpha}, \hat{\psi}) \}^{-1} \{ P_n m_C(\hat{\alpha}, \hat{\psi})m_C(\hat{\alpha}, \hat{\psi})^T \} \{ P_n \dot{m}_C(\hat{\alpha}, \hat{\psi}) \}^{-1T}$. Furthermore, when $g(H_t)^T \alpha$ is a correct model for $\mu(H_t)$ in the sense that there exists $\alpha^*$ such that $g(H_t)^T \alpha^* = \log E(Y_{t+1} \mid H_t, A_t = 0)$, $\hat{\psi}$ achieves the semiparametric efficiency bound of the semiparametric model defined in Theorem 1.

**Remark 1.** $m_C(\alpha, \psi)$ is robust in the sense that the resulting estimator $\hat{\psi}$ is consistent even if $\exp\{g(H_t)^T \alpha\}$ is a misspecified model for $E(Y_{t+1} \mid H_t, I_t = 1, A_t = 0)$. This robustness results from the orthogonality between the so-called “blipped-down outcome” (Robins 1997), $\exp\{-A_t f(H_t)^T \psi^*\} Y_{t+1}$, and the centered action, $A_t - p_t(H_t)$: $E[\exp\{-A_t f(H_t)^T \psi^*\} Y_{t+1} \{ A_t - p_t(H_t) \} \mid H_t] = 0$, which follows from an important property of the blipped-down outcome: $E[\exp\{-A_t f(H_t)^T \psi^*\} Y_{t+1} \mid H_t, A_t] = E\{Y_{t+1}(\tilde{A}_{t-1}, 0) \mid H_t, A_t\}$. This property plays a key role in the robustness of both the estimator in Theorem 2 and the estimator we develop in Section 5.
5 Estimator for the marginal excursion effect

Now we focus on estimation of $\beta_M(t, S_t)$ where $S_t$ is an arbitrary subset of $H_t$. Suppose $\Delta \geq 1$ is a positive integer. Recall that

$$\beta_M(t, S_t) = \log \frac{E \left[ \prod_{t=0}^{T+1} \frac{1(A_t=0)}{1-p_t(H_t)} Y_{t, \Delta} \mid A_t = 1, H_t, I_t = 1 \right]}{E \left[ \prod_{t=0}^{T+1} \frac{1(A_t=0)}{1-p_t(H_t)} Y_{t, \Delta} \mid A_t = 0, H_t, I_t = 1 \right]}.$$ 

We make a parametric assumption on $\beta_M(t, S_t)$. Suppose that for $1 \leq t \leq T$,

$$\beta_M(t, S_t) = S_t^T \beta,$$  

(9)

for some $p$-dimensional parameter $\beta$. Note this model allows for time-dependent effects; $S_t$ could include a vector of basis functions of $t$. The estimation method described below readily generalizes to situations where the parametric model has a known functional form that may be nonlinear; the use of a linear model here enhances presentation clarity.

We propose to use a marginal generalization of the estimating function \[5\] to estimate $\beta$. In particular, the estimating function is

$$m_M(\alpha, \beta) = \sum_{t=1}^{T+\Delta-1} I_t e^{-A_t S_t^T \beta} \left\{ Y_{t, \Delta} - e^{g(H_t)^T \alpha + A_t S_t^T \beta} \right\} J_t \left[ \begin{array}{c} g(H_t) \\ \{A_t - \tilde{p}_t(S_t)\} S_t \end{array} \right].$$  

(10)

where $\exp\{g(H_t)^T \alpha\}$ is a working model for $E\{Y_{t, \Delta}(\tilde{A}_{t-1}, 0, 0) \mid H_t, I_t = 1, A_t = 0\}$ as before. Because the model is now on the marginal effect, we apply a weighting and centering technique similar to \[8\] Boruvka et al. (2018). The weight at time $t$ is

$$J_t = \left\{ \frac{\tilde{p}_t(S_t)}{p_t(H_t)} \right\}^{A_t} \left\{ \frac{1 - \tilde{p}_t(S_t)}{1 - p_t(H_t)} \right\}^{1 - A_t} \times \prod_{j=t+1}^{T+\Delta-1} \frac{1(A_j = 0)}{1 - p_j(H_j)},$$  

(11)

where $\tilde{p}_t(S_t) \in (0, 1)$ is arbitrary as long as it depends on $H_t$ only through $S_t$. The product, $\prod_{j=t+1}^{T+\Delta-1} 1(A_j = 0)/(1 - p_j(H_j))$, is standard inverse probability weighting for settings with $\Delta > 1$. The ratio of probabilities, $\{\tilde{p}_t(S_t)/p_t(H_t)\}^{A_t}[\{(1 - \tilde{p}_t(S_t))/(1 - p_t(H_t))\}^{1 - A_t}$, can be viewed as a change of probability: intuitively, the ratio transforms the data distribution in which $A_t$ is randomized with probability $p_t(H_t)$ to a distribution acting as if $A_t$ were
randomized with probability $\tilde{p}_t(S_t)$. We thus center $A_t$ with $\tilde{p}_t(S_t)$; this centering results in orthogonality between the estimation of $\beta$ and the estimation of the nuisance parameter, $\alpha$. The weighting and centering, together with the factor $\exp(-A_tS_t^T\beta)$, makes the resulting estimator for $\beta$ consistent even when the working model $\exp\{g(H_t)^T\alpha\}$ is misspecified.

In Appendix C we prove the following result.

**Theorem 3.** Suppose (9) and Assumptions 1, 2 and 3 hold. Suppose $\beta^*$ is the value of $\beta$ corresponding to the data generating distribution, $P_0$. Let $\hat{\alpha}$ be a solution to $P_n\hat{m}_M(\hat{\alpha},\hat{\beta}) = 0$. Under regularity conditions, $\sqrt{n}(\hat{\beta} - \beta^*)$ is asymptotically normal with mean zero and covariance matrix $\Sigma_M$. A consistent estimator for $\Sigma_M$ is the lower block diagonal $(p \times p)$ entry of the matrix $\left\{P_n\hat{m}_M(\hat{\alpha},\hat{\beta})\right\}^{-1}\{P_n\hat{m}_M(\hat{\alpha},\hat{\beta})m_M(\hat{\alpha},\hat{\beta})\} \{P_n\hat{m}_M(\hat{\alpha},\hat{\beta})\}^{-1T}$.

**Remark 2.** The consistency of $\hat{\beta}$ does not require the working model $\exp\{g(H_t)^T\alpha\}$ to be correctly specified. This robustness property is desirable because $H_t$ can be high dimensional in an MRT (with the total number of time points, $T$, being hundreds or even thousands), which makes it difficult to model $E\{Y_{t}\Delta(\bar{A}_{t-1},0,0) \mid H_t,I_t=1,A_t=0\}$ correctly.

**Remark 3.** Under the assumptions in Theorem 3, the choice of $\tilde{p}_t(S_t)$ doesn’t affect the consistency of $\hat{\beta}$ as long as it depends at most on $S_t$ and it lies in $(0,1)$. When the parametric model for $\beta_M(t,S_t)$ in (9) is misspecified, $\tilde{p}_t(S_t)$ determines the probability limit of $\hat{\beta}$. For example, when $\Delta=1$ and $S_t=\emptyset$, $\hat{\beta}$ converges in probability to

$$
\beta' = \log \frac{\sum_{t=1}^{T} E\{E(Y_{t+1} \mid H_t,A_t=1) \mid I_t=1\}}{\sum_{t=1}^{T} E\{E(Y_{t+1} \mid H_t,A_t=0) \mid I_t=1\}},
$$

which further simplifies to $\log\{\sum_{t=1}^{T} E(Y_{t+1} \mid I_t=1,A_t=1) / \sum_{t=1}^{T} E(Y_{t+1} \mid I_t=1,A_t=0)\}$ if the randomization probability $p_t(H_t)$ is constant. For general $\Delta$ and $S_t$, the form of the probability limit of $\hat{\beta}$, $\beta'$, is provided in Appendix D.
6 Simulation

6.1 Overview

Here we focus on the causal excursion effect with $\Delta = 1$, and we conduct two simulation studies to evaluate the proposed estimator of the marginal excursion effect ("EMEE") in Section 5 and the semiparametric, locally efficient estimator of the conditional effect ("ECE") described in Section 4.

Because the sandwich estimator for the variance of EMEE in Theorem 3 can be anti-conservative when the sample size is small, we adopt the small sample correction technique in [Mancl & DeRouen 2001] to modify the term $\mathbb{P}_n m_M(\hat{\alpha}, \hat{\beta})^{\otimes 2}$ in the variance estimator. In particular, we pre-multiply the vector of each individual’s residual, $(Y_{t+1} - \exp\{g(H_t)^T \hat{\alpha} + A_t S_t^T \hat{\beta}\}: 1 \leq t \leq T)$, by the inverse of the identity matrix minus the leverage for this individual. Also, as in [Liao et al. 2016], we use critical values from a $t$ distribution. In particular, for a known $p$-dimensional vector $c$, to test the null hypothesis $c^T \beta = 0$ or to form two-sided confidence intervals, we use the critical value $t_{n-p-q}(1 - \xi/2)$, where $p,q$ are the dimensions of $\beta, \alpha$, respectively, and $\xi$ is the significance level. A similar correction is also applied to the variance estimator of ECE in Theorem 2.

The numerical algorithm that solves $\mathbb{P}_n m_C(\alpha, \psi) = 0$ can be unstable when the denominator in $\tilde{K}_t$ gets close to 0. This is because $\exp\{g(H_t)^T \alpha\}$ and $\exp\{g(H_t)^T \alpha + f(H_t)^T \psi\}$ are not constrained within $(0,1)$. In our implementation of ECE, to improve the numerical stability we replace $\tilde{K}_t$ in (8) by

$$e^{f(H_t)^T \psi}
\frac{e^{f(H_t)^T \psi}[1 - \max\{e^{g(H_t)^T \alpha}, \lambda\}]p_t(H_t) + [1 - \max\{e^{g(H_t)^T \alpha + f(H_t)^T \psi}, \lambda\}]\{1 - p_t(H_t)\}}\{1 - \max\{e^{g(H_t)^T \alpha}, \lambda\}\},$$

with the thresholding parameter value $\lambda = 0.95$.

Throughout the simulations, we assume that all individuals are available at all time points, and we omit $I_t = 1$ in writing conditional expectations.

R code [R Core Team 2018] to reproduce the simulation results can be downloaded at https://github.com/tqian/binary-outcome-mrt.
6.2 Simulation on consistency

Here we illustrate that the use of ECE to estimate \( \beta_0 \), by setting \( f(H_t) = 1 \) in (8), can result in an inconsistent estimator yet EMEE with \( S_t = 1 \) yields a consistent estimator. For comparison, we also include the generalized estimating equation (GEE) estimator for binary outcome with log link in the simulation, as GEE is widely used in analyzing mHealth data \cite{Schwartz2007, Bolger2013}. We use independence ("GEE.ind") and exchangeable ("GEE.exch") as working correlation structures for GEE. In all cases the working model \( g(H_t)^T \alpha \) will be misspecified.

The generative model is as follows. The time-varying covariate, \( Z_t \), is independent of all variables observed before \( Z_t \), and it takes three values 0, 1, 2 with equal probability. The randomization probability is constant with \( p_t(H_t) = 0.2 \). The outcome \( Y_{t+1} \) is generated from a Bernoulli distribution with

\[
E(Y_{t+1} | H_t, A_t) = \{0.21(Z_t = 0) + 0.51(Z_t = 1) + 0.41(Z_t = 2)\}e^{A_t(0.1 + 0.3Z_t)}.
\]

We are interested in estimating the fully marginal excursion effect, which equals

\[
\beta_0 = \log \frac{E\{E(Y_{t+1} | H_t, A_t = 1)\}}{E\{E(Y_{t+1} | H_t, A_t = 0)\}} = 0.477.
\]

Throughout we use working model \( g(H_t)^T \alpha = \alpha_0 + \alpha_1 Z_t \), which is misspecified, for all estimators.

The simulation result is given in Table 1; the total number of time points is \( T = 30 \) for each individual. The bias, standard deviation (SD), root mean squared error (RMSE), 95% confidence interval coverage probability before small sample correction (CP (unadj)) and after small sample correction (CP (adj)) are all computed based on 1,000 replicates. As expected, EMEE consistently estimates \( \beta_0 \), and the incorrect use of ECE results in an inconsistent estimator for \( \beta_0 \). The consistency of GEE generally requires the working model \( g(H_t)^T \alpha \) to be correct; in other words, it does not have the robustness property as EMEE. The result shows that both GEE.ind and GEE.exch are inconsistent. We also see that small sample correction helps to improve the confidence interval coverage for EMEE.
Additional simulation results with this generative model where we set $S_t = Z_t$ in EMEE and $f(H_t) = Z_t$ in ECE are given in Appendix E.1.

### 6.3 Simulation on efficiency

Here we focus on the relative efficiency between EMEE and ECE when the marginal excursion effect equals the conditional effect, in which case both estimators are consistent. The relative efficiency is defined as $\text{Var}(\text{EMEE})/\text{Var}(\text{ECE})$, and a quantity larger than 1 means that ECE is more efficient than EMEE. We shall see that if one had adequate data so as to consistently estimate the potentially complex, high dimensional $E(Y_{t+1} \mid H_t, A_t = 0)$ and the marginal excursion effect equals the conditional effect, then ECE can be more efficient than EMEE.

We use the following generative model. The time-varying covariate $Z_t$ is generated from an autoregressive process: $Z_t = 0.5 Z_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0,1)$ is independent of all the variables observed prior to $Z_t$. The randomization probability is given by $p_t(H_t) = \min[0.8, \max\{0.2, \expit(\eta Z_t)\}]$, where $\expit(x) = \{1 + \exp(-x)\}^{-1}$. The proximal outcome $Y_{t+1}$ depends on $(A_{t-1}, Y_t, Z_t, A_t)$ through

$$E(Y_{t+1} \mid H_t, A_t) = q(Z_t, Y_t, A_{t-1}; \gamma)e^{\beta_0 A_t}.$$  

We consider two different $q(Z_t, Y_t, A_{t-1}; \gamma)$:

$$q_{\text{exp}}(Z_t, Y_t, A_{t-1}; \gamma) = \min\{0.8, \max\{0.1, \exp(-0.4 + \gamma(Z_t - 3) + 0.2 Y_t + 0.2 A_{t-1})\}\},$$

and

$$q_{\text{expit}}(Z_t, Y_t, A_{t-1}; \gamma) = \min\{0.8, \max\{0.1, \expit(-0.5 + \gamma Z_t + 0.2 Y_t + 0.2 A_{t-1})\}\}.$$

We fix $\beta_0 = 0.1$.

We consider estimation of $\beta_0$ under the class of generative models with $\eta = -0.5, 0.5$ and $\gamma = 0.1, 0.3, 0.5$. The parameter $\eta$ encodes how the randomization probability depends on $Z_t$, and $\gamma$ encodes the impact of $Z_t$ on the proximal outcome $Y_{t+1}$. We set $f(H_t) = 1$ and $S_t = 1$ in the analysis models of ECE and EMEE, respectively. Because in the generative model $\beta_C(t, H_t) = \beta_M(t, S_t) = \beta_0$, both estimators are consistent for $\beta_0$. We use the working model $g(H_t)^T \alpha = \alpha_0 + \alpha_1 Z_t$, which is misspecified, for both estimators.
Table 1: Performance of EMEE, ECE, GEE.ind and GEE.exch for the marginal excursion effect $\beta_0$.

| Estimator   | Sample size | Bias  | SD    | RMSE  | CP (unadj) | CP (adj) |
|-------------|-------------|-------|-------|-------|------------|----------|
| EMEE        | 30          | 0.000 | 0.077 | 0.077 | 0.93       | 0.94     |
|             | 50          | 0.001 | 0.057 | 0.057 | 0.94       | 0.95     |
|             | 100         | 0.000 | 0.041 | 0.041 | 0.95       | 0.95     |
| ECE         | 30          | 0.048 | 0.075 | 0.089 | 0.85       | 0.88     |
|             | 50          | 0.049 | 0.055 | 0.074 | 0.84       | 0.85     |
|             | 100         | 0.048 | 0.040 | 0.063 | 0.75       | 0.76     |
| GEE.ind     | 30          | 0.041 | 0.073 | 0.084 | 0.88       | 0.89     |
|             | 50          | 0.042 | 0.054 | 0.069 | 0.86       | 0.87     |
|             | 100         | 0.041 | 0.039 | 0.056 | 0.80       | 0.81     |
| GEE.exch    | 30          | 0.041 | 0.073 | 0.084 | 0.87       | 0.89     |
|             | 50          | 0.042 | 0.054 | 0.069 | 0.86       | 0.88     |
|             | 100         | 0.041 | 0.039 | 0.056 | 0.80       | 0.81     |

* EMEE: the estimator of the marginal excursion effect proposed in Section 5. ECE: the semiparametric, locally efficient estimator of the conditional effect described in Section 4. GEE.ind: GEE with independence working correlation structure. GEE.exch: GEE with exchangeable working correlation structure. SD: standard deviation. RMSE: root mean squared error. CP: 95% confidence interval coverage probability, before (unadj) and after (adj) small sample correction. Boldface indicates when Bias or CP are significantly different, at the 5% level, from 0 or 0.95, respectively. Sample size refers to the number of individuals in each simulated trial.
Figure 1 shows the relative efficiency under different combinations of \((\eta, \gamma)\) and the two choices of \(q(\cdot)\) calculated from 1,000 replicates. The sample size is 50, and the total number of time points for each individual is 20. The relative efficiency between the two estimators ranges between 1.11 and 1.00, indicating that there could be slight efficiency gain by using ECE when both estimators are consistent.

Figure 1: Relative efficiency between ECE and EMEE, defined as \(\frac{\text{Var(EMEE)}}{\text{Var(ECE)}}\).

We have also tried other generative models (simulation results not reported here), and small efficiency gain from ECE is observed for most of the cases. Among all the generative models we tried, the only setting where we observe a substantial relative efficiency (\(\sim 1.5\)) is a generative model where the true \(E(Y_{t+1} | H_t, A_t = 0)\) is constant, so that the working model \(g(H_t)^T \alpha\) is always correctly specified. Thus if one had adequate data so as to consistently estimate the potentially complex, high dimensional \(E(Y_{t+1} | H_t, A_t = 0)\) and one felt confident that there are no covariates in \(H_t\) that interact with treatment, then it could be worthwhile to focus on ECE. For completeness, we include the simulation result under this setting in Appendix E.2.
7 Application

BariFit is a 16-week MRT conducted in 2017 by Kaiser Permanente, which aimed to promote weight maintenance for those who went through Bariatric surgery [Ridpath 2017]. In this section, we assess the effect of the food track reminder on individuals’ food log completion rate using estimation methods proposed in this paper. The data set contains 45 participants. The food track reminder was randomly delivered to each participant with probability 0.5 every morning as a text message. Because of the form of the intervention, all participants were available for this intervention throughout the study. The binary proximal outcome, food log completion, is coded as 1 for a day if a participant logged > 0 calories in the Fitbit app on that day.

We used EMEE and ECE for estimating the marginal excursion effect of the food tracker reminder on food log completion, by setting $S_t = 1$ in EMEE and $f(H_t) = 1$ in ECE. We included the day in study (coded as 0,1,...,111), gender, and lag-1 outcome (whether the individual completed food log on the previous day) in the control variables $g(H_t)$. The estimated marginal excursion effect is shown in Table 2 along with the estimated standard error with small sample correction, 95% confidence interval, and $p$-value. Both estimators give qualitatively similar results that no marginal excursion effect is detectable from the data.

The result indicates that no effect of the food track reminder is detectable from the data. There are two possible reasons for the result, which are interrelated. One is an insufficient sample size; this study was not sized to test this particular hypothesis. The other reason is that the true effect may be small or there may be no effect. These findings may inform the next iteration of BariFit study in the following ways. If the researchers want to improve the effectiveness of the food track reminder, they may consider implementing it as a notification with a smartphone app. The current reminder is sent as text message, which cannot be tailored to the user’s current context such as location or weather. Such tailoring may improve effectiveness of the reminder. Alternatively, if the researchers no longer wish to investigate the proximal effect of the food track reminder, they may choose
Table 2: Estimated marginal excursion effect of food track reminder from BariFit data.

| Estimator | Estimate | SE   | 95% CI       | p-value |
|-----------|----------|------|--------------|---------|
| EMEE      | 0.014    | 0.021| (-0.028, 0.056) | 0.50    |
| ECE       | 0.011    | 0.014| (-0.017, 0.039) | 0.44    |

* EMEE: the estimator of the marginal excursion effect proposed in Section 5. ECE: the semiparametric, locally efficient estimator of the conditional effect described in Section 4. SE: standard error. 95% CI: 95% confidence interval. SE, 95% CI and p-value are based on small sample correction described in Section 6.1.

not to randomize it in the next iteration of BariFit. This might be done by either combining the food track reminder with other messages that will be sent in the morning, or to remove the food track reminder completely from the intervention. This can help to reduce the burden of the mHealth intervention on the individual.

8 Discussion

The causal excursion effect defined in this paper is different from the majority of the literature on causal inference in longitudinal setting (Robins 1994, 2000; Van der Laan & Robins 2003). Rather than a contrast of the expected outcome under two fixed treatment histories, the causal excursion effect is a contrast of two “excursions” into the future. The past treatments in the two excursions are considered random (with randomization probability determined by the study design), and are integrated over in the marginalization. We believe the causal excursion effect is a suitable estimand for the primary and secondary analyses in MRT, mainly because it aligns with the domain scientists’ understanding of the estimand in experimental studies in general, which are marginal in nature. Neugebauer et al. (2007) considered a related marginalization idea in MSMs, and argued that such marginalization should be preferred due to its computational tractability, relevance
to public health research, and statistical power.

We treated the model for the proximal outcome under no treatment, \( E\{Y_{t,\Delta}(\bar{A}_{t-1},0,0) \mid H_{t}, I_{t} = 1, A_{t} = 0\} \), as a nuisance parameter, and used a working model \( \exp\{g(H_t)^T \alpha\} \) for this nuisance parameter to reduce noise. In a series of works considering modeling of the treatment effect on a binary outcome in both cross-sectional (Richardson et al. 2017) and longitudinal settings (Wang et al. 2017), those authors propose to instead use log odds-product as the nuisance parameter. This way the nuisance parameter is no longer constrained by the treatment effect model. (As discussed by these authors, the valid range of \( E\{Y_{t,\Delta}(\bar{A}_{t-1},0,0) \mid H_{t}, I_{t} = 1, A_{t} = 0\} \) is constrained by the treatment effect model, because \( E\{Y_{t,\Delta}(\bar{A}_{t-1},1,0) \mid H_{t}, I_{t} = 1, A_{t} = 1\} \) must be within \([0,1]\).) We agree that this congeniality issue is critical when prediction is the goal as the nuisance part of the model would then be of interest, or when the estimation method for the parameters in the treatment effect depends on the correct specification of the nuisance part of the model to be consistent. In the analysis of MRT data, however, the nuisance part of the model is of minimal interest, and more importantly consistency of the estimation methods developed in this paper do not depend on the correct specification of the nuisance part of the model. Therefore, since the purpose of modeling the nuisance parameter is to reduce noise, we choose to treat \( E\{Y_{t,\Delta}(\bar{A}_{t-1},0,0) \mid H_{t}, A_{t} = 0\} \) as a nuisance parameter, because the interpretability makes it easier for domain scientists to model. The estimated probability exceeding \([0,1]\) can sometimes cause numerical instability in the semiparametric, locally efficient estimator described in Section 4 and we addressed this by using the modified weights in (12).

There are a few directions for future research. First, we have assumed binary treatment in the paper. Extension to treatment with multiple levels could involve modeling the treatment effect (defined as contrast to a reference level) as a function of the treatment level. Second, we have focused on estimating the marginal excursion effect. An interesting extension is to introduce random effects to the excursion effect and allow person-specific predictions. With random effects it would be nontrivial to deal with both the nonlinear link function as well as the marginalization. Third, since there are numerous potential variables that can be included in \( g(H_t) \) for noise reduction, one could, because of the
high dimensionality of $H_t$, consider penalized methods for model selection in building the working model $g(H_t)^T \alpha$.

Finally, we note that we used an preliminary version of the estimator for the marginal excursion effect in analyzing the effect of push notification on user engagement in Bidargaddi et al. (2018).

Acknowledgement

Research reported in this paper was supported by National Institute on Alcohol Abuse and Alcoholism (NIAAA) of the National Institutes of Health under award number R01AA23187, National Institute on Drug Abuse (NIDA) of the National Institutes of Health under award number P50DA039838, National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under award number U54EB020404, National Cancer Institute (NCI) of the National Institutes of Health under award number U01CA229437, and National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health under award number R01HL125440. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Akhiezer, N. I. & Glazman, I. M. (2013), Theory of linear operators in Hilbert space, Courier Corporation.

Bidargaddi, N., Almirall, D., Murphy, S., Nahum-Shani, I., Kovalcik, M., Pituch, T., Maaieh, H. & Strecher, V. (2018), ‘To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app’, JMIR mHealth and uHealth 6(11).
Bolger, N. & Laurenceau, J.-P. (2013), *Intensive longitudinal methods: An introduction to diary and experience sampling research*, Guilford Press.

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.

Dempsey, W., Liao, P., Klasnja, P., Nahum-Shani, I. & Murphy, S. A. (2015), ‘Randomised trials for the fitbit generation’, *Significance* **12**(6), 20–23.

Dempsey, W., Liao, P., Kumar, S. & Murphy, S. A. (2017), ‘The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments’, *arXiv preprint arXiv:1711.03587*.

Free, C., Phillips, G., Galli, L., Watson, L., Felix, L., Edwards, P., Patel, V. & Haines, A. (2013), ‘The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review’, *PLoS medicine* **10**(1), e1001362.

Greenland, S. (1987), ‘Interpretation and choice of effect measures in epidemiologic analyses’, *American journal of epidemiology* **125**(5), 761–768.

Heron, K. E. & Smyth, J. M. (2010), ‘Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments’, *British journal of health psychology* **15**(1), 1–39.

Hong, G. & Raudenbush, S. W. (2006), ‘Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data’, *Journal of the American Statistical Association* **101**(475), 901–910.

Hudgens, M. G. & Halloran, M. E. (2008), ‘Toward causal inference with interference’, *Journal of the American Statistical Association* **103**(482), 832–842.
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.

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B. & Murphy, S. A. (2018), ‘Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of heartsteps’, *Annals of Behavioral Medicine*.

Kramer, J.-N., Künzler, F., Mishra, V., Presset, B., Kotz, D., Smith, S., Scholz, U. & Kowatsch, T. (2019), ‘Investigating intervention components and exploring states of receptivity for a smartphone app to promote physical activity: Protocol of a microrandomized trial’, *JMIR research protocols* **8**(1), e11540.

Laird, N. M. & Ware, J. H. (1982), ‘Random-effects models for longitudinal data’, *Biometrics* **38**(4), 963–974.

Law, L. M., Edirisinghe, N. & Wason, J. M. (2016), ‘Use of an embedded, micro-randomised trial to investigate non-compliance in telehealth interventions’, *Clinical Trials* **13**(4), 417–424.

Liang, K.-Y. & Zeger, S. L. (1986), ‘Longitudinal data analysis using generalized linear models’, *Biometrika* **73**(1), 13–22.

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.

Lumley, T., Kronmal, R. & Ma, S. (2006), ‘Relative risk regression in medical research: models, contrasts, estimators, and algorithms’, *Unpublished paper, available at https://biostats.bepress.com/uwbiostat/paper293/.*

Mancl, L. A. & DeRouen, T. A. (2001), ‘A covariance estimator for gee with improved small-sample properties’, *Biometrics* **57**(1), 126–134.
Neugebauer, R., van der Laan, M. J., Joffe, M. M. & Tager, I. B. (2007), ‘Causal inference in longitudinal studies with history-restricted marginal structural models’, *Electronic journal of statistics* 1, 119.

Newey, W. K. (1990), ‘Semiparametric efficiency bounds’, *Journal of applied econometrics* 5(2), 99–135.

Pepe, M. S. & Anderson, G. L. (1994), ‘A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data’, *Communications in Statistics-Simulation and Computation* 23(4), 939–951.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

URL: https://www.R-project.org/

Rabbi, M., Kotov, M. P., Cunningham, R., Bonar, E. E., Nahum-Shani, I., Klasnja, P., Walton, M. & Murphy, S. (2018), ‘Toward increasing engagement in substance use data collection: development of the substance abuse research assistant app and protocol for a microrandomized trial using adolescents and emerging adults’, *JMIR research protocols* 7(7).

Richardson, T. S., Robins, J. M. & Wang, L. (2017), ‘On modeling and estimation for the relative risk and risk difference’, *Journal of the American Statistical Association* 112(519), 1121–1130.

Ridpath, J. (2017), ‘How can we use technology to support patients after bariatric surgery?’, https://www.kpwashingtonresearch.org/news-and-events/recent-news/news-2017/how-can-we-use-technology-support-patients-after-bariatric-surgery/. Accessed: 2019-02-15.

Robins, J. (1986), ‘A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect’, *Mathematical modelling* 7(9-12), 1393–1512.
Robins, J. M. (1994), ‘Correcting for non-compliance in randomized trials using structural
nested mean models’, *Communications in Statistics-Theory and methods* **23**(8), 2379–
2412.

Robins, J. M. (1997), Causal inference from complex longitudinal data, *in* ‘Latent variable
modeling and applications to causality’, Springer, pp. 69–117.

Robins, J. M. (2000), Marginal structural models versus structural nested models as tools
for causal inference, *in* ‘Statistical models in epidemiology, the environment, and clinical
trials’, Springer, pp. 95–133.

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

Rubin, D. B. (1974), ‘Estimating causal effects of treatments in randomized and nonran-
domized studies.’, *Journal of educational Psychology* **66**(5), 688.

Schildcrout, J. S. & Heagerty, P. J. (2005), ‘Regression analysis of longitudinal binary
data with time-dependent environmental covariates: bias and efficiency’, *Biostatistics*
**6**(4), 633–652.

Schwartz, J. E. & Stone, A. A. (2007), The analysis of real-time momentary data: A
practical guide, *in* ‘The science of real-time data capture: Self-reports in health research’,
Oxford University Press New York, NY, pp. 76–113.

Spring, B. (2017), ‘Sense2stop: Mobile sensor data to knowledge. In: ClinicalTrials.gov
[Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2019 May
10]. Available at https://clinicaltrials.gov/ct2/show/nct03184389’.

Tate, D. & Valle, G. (2019), ‘Precision public health: Enhancing connections to develop
just-in-time adaptive intervention strategies (nudge). In: ClinicalTrials.gov [Internet].
Bethesda (MD): National Library of Medicine (US). 2000- [cited 2019 March 23]. Avail-
able at https://clinicaltrials.gov/ct2/show/nct03836391’.
Tsiatis, A. (2007), *Semiparametric theory and missing data*, Springer Science & Business Media.

Van der Laan, M. J. & Robins, J. M. (2003), *Unified methods for censored longitudinal data and causality*, Springer Science & Business Media.

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

Wang, L., Richardson, T. S. & Robins, J. M. (2017), ‘Congenial causal inference with binary structural nested mean models’, *arXiv preprint arXiv:1709.08281*. 
Appendix

A Proof of identifiability result (4)

Lemma A.1. For any $1 \leq k \leq \Delta$, we have

$$
E\{Y_{t,\Delta}(\bar{A}_{t-1},a,\bar{0}) \mid H_t, A_t = a, I_t = 1\}
= E\left\{ \prod_{j=t+1}^{t+k-1} \frac{1(A_j = 0)}{1 - p_j(H_j)} Y_{t,\Delta}(\bar{A}_{t-1},a,\bar{0}) \mid A_t = a, H_t, I_t = 1 \right\}, \quad (A.1)
$$

of Lemma A.1. For $k = 1$, (A.1) holds because we defined $\prod_{j=t+1}^{t+k-1} \frac{1(A_j = 0)}{1 - p_j(H_j)} = 1$. In the following we assume $\Delta \geq 2$, and we prove the lemma by induction on $k = 1, \ldots, \Delta$.

Suppose (A.1) holds for $k = k_0$ for some $1 \leq k_0 \leq \Delta - 1$. Denote by $\zeta = \prod_{j=t+1}^{t+k_0-1} \frac{1(A_j = 0)}{1 - p_j(H_j)} Y_{t,\Delta}(\bar{A}_{t-1},a,\bar{0})$. We have

$$
E(\zeta \mid H_{t+k_0}, A_t = a, I_t = 1)
= E(\zeta \mid H_{t+k_0}, A_t = a, I_t = 1) E\left\{ \prod_{j=t+1}^{t+k_0-1} \frac{1(A_j = 0)}{1 - p_j(H_j)} Y_{t,\Delta}(\bar{A}_{t-1},a,\bar{0}) \mid H_{t+k_0}, A_t = a, I_t = 1 \right\}
= E\left\{ \zeta \times \frac{1(A_{t+k_0} = 0)}{1 - p_{t+k_0}(H_{t+k_0})} Y_{t,\Delta}(\bar{A}_{t-1},a,\bar{0}) \mid H_{t+k_0}, A_t = a, I_t = 1 \right\}
= E\left\{ \prod_{j=t+1}^{t+k_0} \frac{1(A_j = 0)}{1 - p_j(H_j)} Y_{t,\Delta}(\bar{A}_{t-1},a,\bar{0}) \mid H_{t+k_0}, A_t = a, I_t = 1 \right\}, \quad (A.2)
$$

where (A.2) follows from sequential ignorability (Assumption 3). Therefore, by the induction hypothesis and the law of iterated expectation we have

$$
E\{Y_{t,\Delta}(\bar{A}_{t-1},a,\bar{0}) \mid H_t, A_t = a, I_t = 1\}
= E(\zeta \mid H_t, A_t = a, I_t = 1)
= E\left\{ \prod_{j=t+1}^{t+k_0-1} \frac{1(A_j = 0)}{1 - p_j(H_j)} Y_{t,\Delta}(\bar{A}_{t-1},a,\bar{0}) \mid H_t, A_t = a, I_t = 1 \right\}, \quad (A.3)
$$

i.e., we showed that (A.1) holds for $k = k_0 + 1$. This completes the proof. □
of identifiability result (4). It suffices to show that under Assumptions 1-3, we have

\[
E\{Y_t, \Delta(\bar{A}_{t-1}, a, \bar{0}) \mid S_t(\bar{A}_{t-1}), I_t(\bar{A}_{t-1}) = 1\}
= E\left[ E\left\{ \prod_{j=t+1}^{t+\Delta-1} 1(A_j = 0) \frac{1}{1 - p_j(H_j)} Y_{t, \Delta} \left| A_t = a, H_t, I_t = 1 \right. \right\} \mid S_t, I_t = 1 \right]. \tag{A.4}
\]

We have the following sequence of equality:

\[
E\{Y_t, \Delta(\bar{A}_{t-1}, a, \bar{0}) \mid S_t(\bar{A}_{t-1}), I_t(\bar{A}_{t-1}) = 1\}
= E[E\{Y_t, \Delta(\bar{A}_{t-1}, a, \bar{0}) \mid H_t(\bar{A}_{t-1}), I_t(\bar{A}_{t-1}) = 1\} \mid S_t(\bar{A}_{t-1}), I_t(\bar{A}_{t-1}) = 1] \tag{A.5}
= E[E\{Y_t, \Delta(\bar{A}_{t-1}, a, \bar{0}) \mid H_t, I_t = 1\} \mid S_t, I_t = 1] \tag{A.6}
= E[E\{Y_t, \Delta(\bar{A}_{t-1}, a, \bar{0}) \mid H_t, A_t = a, I_t = 1\} \mid S_t, I_t = 1] \tag{A.7}
= E \left[ E\left\{ \prod_{j=t+1}^{t+\Delta-1} 1(A_j = 0) \frac{1}{1 - p_j(H_j)} Y_{t, \Delta} \left| A_t = a, H_t, I_t = 1 \right. \right\} \mid S_t, I_t = 1 \right], \tag{A.8}
\]

where (A.5) follows from the law of iterated expectation, (A.6) follows from consistency (Assumption 1), (A.7) follows from sequential ignorability (Assumption 3), and (A.8) follows from Lemma A.1. This completes the proof.

B Proof of Theorem 2

To establish Theorem 2 we assume the following regularity conditions.

Assumption B.1. Suppose \((\alpha, \psi) \in \Theta\), where \(\Theta\) is a compact subset of a Euclidean space. Suppose there exists unique \((\alpha', \psi') \in \Theta\) such that \(E\{m_C(\alpha', \psi')\} = 0\).

Assumption B.2. Suppose \(f(H_t)\) and \(g(H_t)\) are bounded for all \(t\).

Lemma B.1. Suppose (6) and Assumptions 1, 2 and 3 hold. Suppose \(\psi^*\) is the value of \(\psi\) corresponding to the data generating distribution, \(P_0\). For an arbitrary \(\alpha\), we have

\[
E[I_t e^{-A_t f(H_t)^T \psi^*} \{Y_{t+1} - e^{g(H_t)^T \alpha + A_t f(H_t)^T \psi^*} \} \bar{K}_t \{A_t - p_t(H_t)\} f(H_t)] = 0. \tag{B.1}
\]
of Lemma B.1. By the law of iterated expectation we have

\[
E[Ite^{-A_t f(H_t)^T}\psi^* \{Y_{t+1} - e^{g(H_t)\alpha + A_t f(H_t)^T}\hat{\psi}\}] K_t\{A_t - \hat{p}_t(H_t)\} f(H_t)
\]

\[
= E(E[Ite^{-A_t f(H_t)^T}\psi^* \{Y_{t+1} - e^{g(H_t)\alpha + A_t f(H_t)^T}\psi^*\}]K_t\{A_t - \hat{p}_t(H_t)\} f(H_t) | H_t)
\]

\[
= E(E[e^{-f(H_t)^T}\psi^* \{Y_{t+1} - e^{g(H_t)\alpha + f(H_t)^T}\psi^*\}]\{1 - \hat{p}_t(H_t)\} | H_t, I_t = 1)K_t f(H_t)
\]

\[
= E[E[\{Y_{t+1} - e^{g(H_t)\alpha}\}p_t(H_t) | H_t, I_t = 1, \alpha = 0, \hat{p}_t(H_t)\]K_t f(H_t)
\]

\[
= E[\{e^{-f(H_t)^T}\psi^* E(Y_{t+1} | H_t, I_t = 1, \alpha = 0) - E(Y_{t+1} | H_t, I_t = 1, \alpha = 0)\}
\]

\[
\times p_t(H_t)\{1 - \hat{p}_t(H_t)\} K_t f(H_t)
\]

= 0,

where the last equality follows from \([\square]\). This completes the proof. \(\square\)

of Theorem 2. Assumption B.1 implies that \((\hat{\alpha}, \hat{\psi})\) converges in probability to \((\alpha', \psi')\), by Theorem 5.9 and Problem 5.27 of Van der Vaart (2000). Because \(m_C(\alpha, \psi)\) is continuously differentiable and hence Lipschitz continuous, Theorem 5.21 of Van der Vaart (2000) implies that \(\sqrt{n}\{(\hat{\alpha}, \hat{\psi}) - (\alpha', \psi')\}\) is asymptotically normal with mean zero and covariance matrix \(\{E\{\hat{m}_C(\alpha', \psi')\}\}^{-1} E\{m_C(\alpha', \psi')m_C(\alpha', \psi')^T\} E\{\hat{m}_C(\alpha', \psi')\}^{-1T}\). By the law of large numbers and Slutsky’s theorem, this covariance matrix can be consistently estimated by \(\{P_n \hat{m}_C(\hat{\alpha}, \hat{\psi})\}^{-1} \{P_n m_C(\hat{\alpha}, \hat{\psi})m_C(\hat{\alpha}, \hat{\psi})^T\} \{P_n \hat{m}_C(\hat{\alpha}, \hat{\psi})\}^{-1T}\). Furthermore, Assumption B.1 and Lemma B.1 imply that \(\psi^* = \psi'\), so we proved the asymptotic normality of \(\hat{\psi}\). When \(g(H_t)^T\alpha\) is a correct model for \(\mu(H_t)\), that \(\hat{\psi}\) attains the semiparametric efficiency bound follows from Theorem 1. This completes the proof. \(\square\)

C Proof of Theorem 3

The proof of Theorem 3 is similar to the proof of Theorem 2. To establish Theorem 3 we assume the following regularity conditions.
**Assumption C.1.** Suppose \((\alpha, \beta) \in \Theta\), where \(\Theta\) is a compact subset of a Euclidean space. Suppose there exists unique \((\alpha', \beta') \in \Theta\) such that \(E\{m_M(\alpha', \beta')\} = 0\).

**Assumption C.2.** Suppose \(S_t, \exp(S_t), g(H_t)\) and \(\exp\{g(H_t)\}\) all have finite fourth moment.

**Lemma C.1.** Suppose \([\square]\) and Assumptions \([\square \square \square] \) hold. Suppose \(\beta^*\) is the value of \(\beta\) corresponding to the data generating distribution, \(P_0\). For an arbitrary \(\alpha\), we have

\[
E[I_t e^{-A_t S_t^T \beta^*} \{Y_{t+1} - e^{g(H_t)^T \alpha + A_t S_t^T \beta^*}\} J_t \{A_t - \tilde{p}_t(S_t)\} S_t] = 0. \tag{C.1}
\]

**of Lemma C.1.** By the law of iterated expectation we have

\[
E[I_t e^{-A_t S_t^T \beta^*} \{Y_{t,\Delta} - e^{g(H_t)^T \alpha + A_t S_t^T \beta^*}\} J_t \{A_t - \tilde{p}_t(S_t)\} S_t] \\
= E(E[I_t e^{-A_t S_t^T \beta^*} \{Y_{t,\Delta} - e^{g(H_t)^T \alpha + A_t S_t^T \beta^*}\} J_t \{A_t - \tilde{p}_t(S_t)\} S_t | H_t) \\
= E(E[e^{-A_t S_t^T \beta^*} \{Y_{t,\Delta} - e^{g(H_t)^T \alpha + A_t S_t^T \beta^*}\} J_t \{A_t - \tilde{p}_t(S_t)\} | H_t, I_t = 1] S_t) \\
= E\left(E\left[e^{-S_t^T \beta^*} \{Y_{t,\Delta} - e^{g(H_t)^T \alpha + S_t^T \beta^*}\} \{1 - \tilde{p}_t(S_t)\} \prod_{j=t+1}^{t+\Delta-1} \frac{1(A_j = 0)}{1 - p_j(H_j)} \left| H_t, I_t = 1, A_t = 1\right] \tilde{p}_t(S_t) S_t \right) \\
- E\left(E\left[\{Y_{t,\Delta} - e^{g(H_t)^T \alpha}\} \tilde{p}_t(S_t) \prod_{j=t+1}^{t+\Delta-1} \frac{1(A_j = 0)}{1 - p_j(H_j)} \left| H_t, I_t = 1, A_t = 0\right] \{1 - \tilde{p}_t(S_t)\} S_t \right) \\
= E\left[\left\{e^{-S_t^T \beta^*} E\left(\prod_{j=t+1}^{t+\Delta-1} \frac{1(A_j = 0)}{1 - p_j(H_j)} Y_{t,\Delta} \left| H_t, I_t = 1, A_t = 1\right) \right) \right\} \tilde{p}_t(S_t) \{1 - \tilde{p}_t(S_t)\} S_t \right] \tag{C.2}
\]

where the last equality follows from \([\square]\). This completes the proof.

**of Theorem C.1**. Assumption C.1 implies that \((\hat{\alpha}, \hat{\beta})\) converges in probability to \((\alpha', \beta')\), by Theorem 5.9 and Problem 5.27 of Van der Vaart (2000). Because \(m_M(\alpha, \beta)\) is continuously differentiable and hence Lipschitz continuous, Theorem 5.21 of Van der Vaart (2000) implies that \(\sqrt{n}\{(\hat{\alpha}, \hat{\beta}) - (\alpha', \beta')\}\) is asymptotically normal with mean zero and covariance matrix \([E\{m_M(\alpha', \beta')\}]^{-1} E\{m_M(\alpha', \beta') m_M(\alpha', \beta')^T\} [E\{m_M(\alpha', \beta')\}]^{-1}\). By the law of large
numbers and Slutsky’s theorem, this covariance matrix can be consistently estimated by \( \{P_n \hat{m}_M(\hat{\alpha}, \hat{\beta})\}^{-1}\{P_n m_M(\hat{\alpha}, \hat{\beta})\} m_M(\hat{\alpha}, \hat{\beta})^T \{P_n \hat{m}_M(\hat{\alpha}, \hat{\beta})\}^{-1} \). Furthermore, Assumption B.1 and Lemma B.1 imply that \( \beta^* = \beta' \). This completes the proof.

\[ \square \]

D Limit of \( \hat{\beta} \) in Remark 3 for general \( \Delta \)

When \( \text{(9)} \) is misspecifed, the limit of \( \hat{\beta} \) is \( \beta' \) that satisfies the following equation:

\[
\sum_{t=1}^{T} E \left[ \left( E \left\{ Y_{t+\Delta} \sum_{j=t+1}^{t+\Delta-1} \frac{1(A_j = 0)}{1-p_j(H_j)} \right| H_t, I_t = 1, A_t = 1 \right) e^{-S_t^{T \beta'}} - E \left( Y_{t+\Delta} \sum_{j=t+1}^{t+\Delta-1} \frac{1(A_j = 0)}{1-p_j(H_j)} \left| H_t, I_t = 1, A_t = 0 \right) \right) \tilde{p}_t(S_t) \{1-\tilde{p}_t(S_t)\} S_t \right] = 0. \tag{D.1}
\]

This is derived in (C.2) in the proof of Lemma C.1.

E Additional simulation results

E.1 Additional simulation on consistency

We use the same generative model as in Section 6.2 and here we set \( S_t = Z_t \) in EMEE and \( f(H_t) = Z_t \) in ECE. Because the generative model implies that

\[
\log \frac{E(Y_{t+1} | H_t, A_t = 1)}{E(Y_{t+1} | H_t, A_t = 0)} = 0.1 + 0.3Z_t,
\]

both EMEE and ECE should consistently estimate \( \beta_0 = 0.1 \) and \( \beta_1 = 0.3 \). We also included GEE.ind and GEE.exch for comparison. Because the working model for \( E(Y_{t+1} | H_t, A_t = 0) \), \( \exp(\alpha_0 + \alpha_1 Z_t) \), is misspecified, both GEE.ind and GEE.exch would be inconsistent for \( \beta_0 \) and \( \beta_1 \).

The simulation result is given in Table E.1, where the total number of time points is 30 for each individual. The bias, standard deviation (SD), root mean squared error (RMSE), 95% confidence interval coverage probability before small sample correction (CP (unadj))
and after small sample correction (CP (adj)) are all computed based on 1,000 replicates. As expected, EMEE and ECE are consistent for $\beta_0$ and $\beta_1$, and GEE.ind and GEE.exch are inconsistent. We also see that ECE is slightly more efficient than EMEE.

E.2 Simulation that presents substantial relative efficiency between ECE and EMEE

Consider the following generative model. The covariate $Z_t$ is exogenous and generated from Uniform $[0,1]$. The treatment indicator $A_t$ is binary with constant randomization probability $\eta$. The outcome is generated from a Bernoulli distribution with mean $E(Y_{t+1} | H_t, A_t) = 0.3\exp\{A_t(\beta_0 + \beta_1 Z_t}\}$. We set $\beta_0 = \log \frac{1}{3}$ and $\beta_1 = 2\log 3$, so that $E(Y_{t+1} | H_t, A_t) \in [0.1, 0.9]$. In the model of ECE we set $f(H_t) = Z_t$, and in the model of EMEE we set $S_t = Z_t$, so that both estimators are consistent for $\beta_0$ and $\beta_1$. The working model for the control part is $\exp\{g_t(H_t)^T \alpha\} = \exp(\alpha_0 + \alpha_1 Z_t)$, which is correctly specified.

We consider the relative efficiency along a 1-dimensional submodel obtained by varying the constant randomization probability $\eta$ from 0.1 to 0.9. Each panel in Figure E.1 represents a specific combination of total time points $T$ and sample size $n$, and the two curves are the relative efficiency between for estimating $\psi_0$ and $\psi_1$ as $\eta$ ranges from 0.1 to 0.9. We considered $T = 10, 30, 50$ and $n = 30, 100, 200$. For certain values of $(n,T)$, the curves are only present for $\eta$ varying in a narrower range than $[0.1, 0.9]$, because under the other settings some simulated data sets are separable and the algorithm for both ECE and EMEE fail to output an estimator for those data sets.

Figure shows that the relative efficiency is always greater than or equal to 1; this is as expected because with a correctly specified model for $E(Y_{t+1} | H_t, A_t = 0)$, ECE achieves the semiparametric efficiency bound asymptotically and is thus more efficient than EMEE. We observe a general pattern across all panels: the relative efficiency is larger when the randomization probability $\eta$ is smaller. The relative efficiency can be as large as over 1.75, when $\eta = 0.1$, $n = 200$, $T = 50$.

Below we provide an intuitive explanation for this pattern by comparing the estimating
Table E.1: Comparison of the three estimators for the treatment effect modification ($S_t = Z_t$), when the treatment effect conditional on the full history is correctly specified.

| Estimator | Sample size | $\beta_0$ | Bias | RMSE | SD  | CP (unadj) | CP (adj) | $\beta_1$ | Bias | RMSE | SD  | CP (unadj) | CP (adj) |
|-----------|-------------|-----------|------|------|-----|------------|----------|-----------|------|------|-----|------------|----------|
| EMEE      | 30          | -0.02     | 0.20 | 0.20 | 0.94| 0.95       | 0.01     | 0.13      | 0.13 | 0.94 | 0.95|
|           | 50          | -0.01     | 0.16 | 0.16 | 0.95| 0.96       | 0.01     | 0.11      | 0.11 | 0.94 | 0.95|
|           | 100         | -0.01     | 0.11 | 0.11 | 0.96| 0.96       | 0.01     | 0.07      | 0.07 | 0.95 | 0.96|
| ECE       | 30          | -0.02     | 0.18 | 0.18 | 0.94| 0.95       | 0.01     | 0.12      | 0.12 | 0.93 | 0.94|
|           | 50          | -0.01     | 0.15 | 0.15 | 0.94| 0.95       | 0.00     | 0.09      | 0.09 | 0.94 | 0.94|
|           | 100         | -0.01     | 0.10 | 0.10 | 0.96| 0.96       | 0.01     | 0.06      | 0.06 | 0.94 | 0.95|
| GEE.ind   | 30          | **0.14**  | 0.21 | 0.15 | **0.82**| 0.85       | **-0.12**| 0.15      | 0.08 | **0.75**| **0.78**|
|           | 50          | **0.15**  | 0.19 | 0.12 | **0.75**| 0.77       | **-0.12**| 0.14      | 0.07 | **0.60**| **0.63**|
|           | 100         | **0.15**  | 0.17 | 0.08 | **0.57**| 0.58       | **-0.12**| 0.13      | 0.05 | **0.33**| **0.34**|
| GEE.exch  | 30          | **0.14**  | 0.21 | 0.15 | **0.82**| 0.85       | **-0.12**| 0.15      | 0.08 | **0.75**| **0.77**|
|           | 50          | **0.15**  | 0.19 | 0.12 | **0.75**| 0.77       | **-0.12**| 0.14      | 0.07 | **0.60**| **0.62**|
|           | 100         | **0.15**  | 0.17 | 0.08 | **0.57**| 0.58       | **-0.12**| 0.13      | 0.05 | **0.33**| **0.34**|

* EMEE: the estimator of the marginal excursion effect proposed in Section 5. ECE: the semi-parametric, locally efficient estimator of the conditional effect described in Section 4. GEE.ind: GEE with independence working correlation structure. GEE.exch: GEE with exchangeable working correlation structure. SD: standard deviation. RMSE: root mean squared error. CP: 95% confidence interval coverage probability, before (unadj) and after (adj) small sample correction. Boldface indicates when Bias or CP are significantly different, at the 5% level, from 0 or 0.95, respectively. Sample size refers to the number of individuals in each simulated trial.
Figure E.1: Relative efficiency between ECE and EMEE in Section E.2. The relative efficiency is defined as $\frac{\text{Var(EMEE)}}{\text{Var(ECE)}}$. 

![Graph showing relative efficiency for different values of T and n, with estimands $\beta_1$ and $\beta_0$.]
equations $m_M$ in (10) and $m_C$ in (8). The summand in $m_C$ has an additional factor compared to $m_M$:

$$
\frac{1}{(1 - e^{g_t(H_t)^T \alpha})p_t + (e^{-S_t^T \beta} - e^{g_t(H_t)^T \alpha})(1 - p_t)}.
$$

(E.1)

The denominator can be equivalently written as $p_t + (1 - p_t)e^{-S_t^T \beta} - e^{g_t(H_t)^T \alpha}$. In the generative model $e^{-S_t^T \beta}$ varies over an interval (roughly $[\frac{1}{3}, 3]$ if one plugs in the true $\beta$), whereas $e^{g_t(H_t)^T \alpha}$ is almost constant (because $E(Y_{t+1} \mid H_t, A_t = 0) = 0.3$ is constant). Therefore, when $p_t$ is close to 1, $m_C$ is close to $m_M$, hence the relative efficiency is close to 1. When $p_t$ is close to 0, the factor (E.1) becomes more variable, making $m_C$ quite different from $m_M$, and hence a larger difference in their variances. Note that this pattern (larger relative efficiency with smaller $p_t$) only holds for this particular generative model, and may not hold in general. The point of this simulation study is that the efficiency gain from using ECE can sometimes be substantial.

**F  Proof of Theorem 1**

**F.1  Overview**

In Section [F.2](#), we present the proof of Theorem 1 based on a general form of the efficient score using semiparametric efficiency theory developed in Section [F.4](#). In Section [F.3](#), we give assumptions that characterize the semiparametric model, and we introduce additional notation that will be used throughout the proof. In Section [F.4](#) we derive the general form of the efficient score using semiparametric efficiency theory. For ease of reading the proofs, the supporting technical lemmas that are used in deriving the general form of the efficient score are presented and proved in Section [G](#). For notation simplicity, this entire section is presented in the case where $I_t = 1$ for all $t$, and we omit the notation $I_t$ throughout.

The techniques in Section [F.4](#) and Section [G](#) follow mostly from Robin’s derivation of the efficient score for structural nested mean models ([Robins](#1994)).
Proof of Theorem 1 from a general form of efficient score

We first present a useful lemma.

Lemma F.1. Suppose that $B, C$ are two random variables, and that $B$ takes binary value $\{0, 1\}$. Suppose $E\{S(B, C) \mid C\} = 0$ for some function $S(B, C)$. Then we have

$$S(B, C) = \{S(1, C) - S(0, C)\} \times \{B - P(B = 1 \mid C)\}. \quad (F.1)$$

of Lemma F.1. Since $B$ takes binary value, we have

$$S(B, C) = S(1, C)B + S(0, C)(1 - B) = \{S(1, C) - S(0, C)\}B + S(0, C). \quad (F.2)$$

We also have

$$E\{S(B, C) \mid C\} = E\{S(B, C) \mid C, B = 1\}P(B = 1 \mid C) + E\{S(B, C) \mid C, B = 0\} \times \{1 - P(B = 1 \mid C)\} = S(1, C)P(B = 1 \mid C) + S(0, C)\{1 - P(B = 1 \mid C)\} = \{S(1, C) - S(0, C)\}P(B = 1 \mid C) + S(0, C). \quad (F.3)$$

Equation (F.3) and $E\{S(B, C) \mid C\} = 0$ imply

$$S(0, C) = -\{S(1, C) - S(0, C)\}P(B = 1 \mid C). \quad (F.4)$$

Combining (F.2) and (F.4) yields (F.1). This completes the proof. \hfill \Box

of Theorem 1. To connect Theorem 1 with the notation used in the rest of this section, let $\psi_0$ be the true value of the parameter $\psi$. Define $V_t = (H_t, A_t)$, $U_{t+1}(\psi) = Y_{t+1}e^{-A_t f(H_t)^T \psi}$, $\dot{U}_{t+1}(\psi) = U_{t+1}(\psi) - E\{U_{t+1}(\psi_0) \mid H_t\}$, and $\hat{W}_t = \text{Var}\{U_{t+1}(\psi_0) \mid V_t\}^{-1}$.

By Lemma F.8, a general form of the efficient score is

$$S_{\text{eff}}(\psi_0) = -\sum_{t=1}^{T} \rho_t \dot{U}_{t+1}(\psi_0), \quad (F.5)$$
where

\[
\rho_t = \left[ E\left\{ \frac{\partial U_{t+1}(\psi_0)}{\partial \psi} \mid V_t \right\} - E\left\{ \frac{\partial U_{t+1}(\psi_0)}{\partial \psi} W_t \mid H_t \right\} E(W_t \mid H_t)^{-1} \right] W_t. \tag{F.6}
\]

Note that \( E(\rho_t \mid H_t) = 0 \); therefore, Lemma F.1 implies

\[
\rho_t = \{ \rho_t(A_t = 1) - \rho_t(A_t = 0) \} \{ A_t - p_t(H_t) \}, \tag{F.7}
\]

where \( \rho_t(A_t = a) \) denotes \( \rho_t \) (as a function of \( H_t \) and \( A_t \)) evaluated at \( A_t = a \). In the following we calculate corresponding terms in the context of Theorem 1.

First, we have

\[
\frac{\partial U_{t+1}(\psi_0)}{\partial \psi} = -U_{t+1}(\psi_0) A_t f(H_t), \tag{F.8}
\]

and hence

\[
E\left\{ \frac{\partial U_{t+1}(\psi_0)}{\partial \psi} \mid H_t, A_t = 1 \right\} = -E\{ U_{t+1}(\psi_0) \mid H_t, A_t = 1 \} f(H_t)
\]

\[
= -E\{ Y_{t+1}(\tilde{A}_{t-1}, 0) \mid H_t \} f(H_t) = -e^{\mu(H_t)} f(H_t), \tag{F.9}
\]

\[
E\left\{ \frac{\partial U_{t+1}(\psi_0)}{\partial \psi} \mid H_t, A_t = 0 \right\} = 0. \tag{F.10}
\]

where the second equality in (F.9) follows from Lemma G.1.

Second, we have

\[
W_t = \text{Var}\{ U_{t+1}(\psi_0) \mid V_t \}^{-1} = \text{Var}(Y_{t+1} \mid V_t)^{-1} e^{2A_t f(H_t)T \psi_0}
\]

\[
= \left[ e^{\mu(H_t)+A_t f(H_t)T \psi_0} \left\{ 1 - e^{\mu(H_t)+A_t f(H_t)T \psi_0} \right\} \right]^{-1} e^{2A_t f(H_t)T \psi_0}
\]

\[
= \frac{e^{A_t f(H_t)T \psi_0}}{e^{\mu(H_t)} \left\{ 1 - e^{\mu(H_t)+f(H_t)T \psi_0} \right\}^2}, \tag{F.11}
\]

\[
W_t(A_t = 1) = \frac{e^{f(H_t)T \psi_0}}{e^{\mu(H_t)} \left\{ 1 - e^{\mu(H_t)+f(H_t)T \psi_0} \right\}^2}, \tag{F.12}
\]

\[
W_t(A_t = 0) = \frac{1}{e^{\mu(H_t)} \left\{ 1 - e^{\mu(H_t)} \right\}^2}, \tag{F.13}
\]

39
and

\[
E(W_t \mid H_t) = E(W_t \mid H_t, A_t = 1)p_t(H_t) + E(W_t \mid H_t, A_t = 0)\{1 - p_t(H_t)\}
\]

\[
= \frac{e^{f(H_t)^T \psi_0}}{e^{\mu(H_t)}\{1 - e^{\mu(H_t) + f(H_t)^T \psi_0}\}}p_t(H_t) + \frac{1}{e^{\mu(H_t)}\{1 - e^{\mu(H_t)}\}}\{1 - p_t(H_t)\}
\]

\[
= \frac{1}{e^{\mu(H_t)}} \times \left\{ e^{f(H_t)^T \psi_0} - e^{\mu(H_t) + f(H_t)^T \psi_0} \right\} p_t(H_t) + \left\{ 1 - e^{\mu(H_t) + f(H_t)^T \psi_0} \right\} \{1 - p_t(H_t)\} . \quad (F.14)
\]

Third, it follows from \( F.8 \) and \( F.11 \) that

\[
E\left\{ \frac{\partial U_t(\psi_0)}{\partial \psi} W_t \mid H_t \right\} = -E\{U_t(\psi_0)A_t f(H_t) W_t \mid H_t\}
\]

\[
= -E\{U_t(\psi_0)W_t \mid H_t, A_t = 1\}p_t(H_t)f(H_t)
\]

\[
= -E\left[ Y_{t+1} e^{-f(H_t)^T \psi_0} \times \frac{e^{f(H_t)^T \psi_0}}{e^{\mu(H_t)}\{1 - e^{\mu(H_t) + f(H_t)^T \psi_0}\}} \mid H_t, A_t = 1\right] p_t(H_t)f(H_t)
\]

\[
= -e^{\mu(H_t)}\frac{e^{f(H_t)^T \psi_0}}{e^{\mu(H_t)}\{1 - e^{\mu(H_t) + f(H_t)^T \psi_0}\}} p_t(H_t)f(H_t) . \quad (F.15)
\]

\[
= -\frac{1}{1 - e^{\mu(H_t) + f(H_t)^T \psi_0}} p_t(H_t)f(H_t) . \quad (F.16)
\]

where \( F.15 \) follows from that \( E[Y_{t+1}\exp\{-f(H_t)^T \psi_0\} \mid H_t, A_t] = E[Y_{t+1}(\tilde{A}_{t-1}, 0) \mid H_t] \), an implication of Lemma \( G.1 \).

Because of \( F.10 \), we have

\[
\rho_t(A_t = 1) - \rho_t(A_t = 0) = E\left\{ \frac{\partial U_{t+1}(\psi_0)}{\partial \psi} \mid H_t, A_t = 1\right\} W_t(A_t = 1)
\]

\[
- E\left\{ \frac{\partial U_{t+1}(\psi_0)}{\partial \psi} W_t \mid H_t\right\} E(W_t \mid H_t)^{-1}\left\{ W_t(A_t = 1) - W_t(A_t = 0) \right\} . \quad (F.17)
\]

By \( F.12 \) and \( F.13 \) we have

\[
W_t(A_t = 1) - W_t(A_t = 0) = \frac{e^{f(H_t)^T \psi_0} - 1}{e^{\mu(H_t)}\{1 - e^{\mu(H_t) + f(H_t)^T \psi_0}\}} . \quad (F.18)
\]

which, combined with \( F.14 \), yields

\[
E(W_t \mid H_t)^{-1}\left\{ W_t(A_t = 1) - W_t(A_t = 0) \right\}
\]

\[
= \frac{e^{f(H_t)^T \psi_0} - 1}{\{e^{f(H_t)^T \psi_0} - e^{\mu(H_t) + f(H_t)^T \psi_0}\} p_t(H_t) + \{1 - e^{\mu(H_t) + f(H_t)^T \psi_0}\}\{1 - p_t(H_t)\}} . \quad (F.19)
\]
Plugging (F.9), (F.12), and (F.19) into (F.17) yields
\[
\rho_t(A_t = 1) - \rho_t(A_t = 0)
\]
\[
= -\frac{e^{f(H_t)^T \psi_0} f(H_t)}{\{e^{f(H_t)^T \psi_0} - e^{\mu(H_t) + f(H_t)^T \psi_0}\} p_t(H_t) + \{1 - e^{\mu(H_t) + f(H_t)^T \psi_0}\} \{1 - p_t(H_t)\}}. 
\]  
(F.20)
Therefore, by plugging (F.20) into (F.7), we have
\[
\rho_t = -K_t \{A_t - p_t(H_t)\} f(H_t). 
\]  
(F.21)
On the other hand, by Lemma G.1 we have
\[
\dot{U}_{t+1}(\psi_0) = U_{t+1}(\psi_0) - E\{U_{t+1}(\psi_0) \mid H_t\} = e^{-A_t f(H_t)^T \psi_0} Y_{t+1} - e^{\mu(H_t)}.
\]  
(F.22)
Plugging (F.21) and (F.22) into (F.5) gives the form of \( S_{\text{eff}}(\psi_0) \). This completes the proof.

**F.3 Assumption and Additional Notation**

In deriving the semiparametric efficient score, we consider the semiparametric model characterized by the following assumptions:

**Assumption F.1.** For all \( 1 \leq t \leq T \), \( E\{Y_{t+1}(\tilde{A}_{t-1},0) \mid H_t, A_t\} = E\{Y_{t+1}(\tilde{A}_{t-1},0) \mid H_t\} \).

**Assumption F.2.** Assume that there exists a function \( \gamma() \) and a true parameter value \( \psi_0 \in \mathbb{R}^p \), such that for any \( 1 \leq t \leq T \),
\[
\log \frac{E\{Y_{t+1}(\tilde{a}_t) \mid \tilde{z}_t, \tilde{a}_t\}}{E\{Y_{t+1}(\tilde{a}_{t-1},0) \mid \tilde{z}_t, \tilde{a}_t\}} = \gamma(t + 1, \tilde{z}_t, \tilde{a}_t; \psi_0).
\]  
(F.23)
In the following, we present additional notation that will be used in the proof. Each will be defined as they appear in the proof. Here we gather the definition of all the terms for ease of reading:

- The longitudinal data is \( L_1, A_1, Y_2, L_2, A_2, Y_3, \ldots, L_T, A_T, Y_{T+1} \), where \( L_t \) is a time-varying covariate, \( A_t \) is the treatment assignment, and \( Y_{t+1} \) is the proximal outcome.
\[ Y_1 = \emptyset, \ L_{T+1} = \emptyset, \ A_{T+1} = \emptyset \]

\[ Z_t = (Y_t, L_t) \]

\[ H_t = (\bar{A}_{t-1}, \bar{Z}_t) \]

\[ V_t = (\bar{A}_t, \bar{Z}_t) = (H_t, A_t) \]

\[ U_{t+1}(\psi) = Y_{t+1} \exp\{-\gamma(t+1, \bar{z}_t, a_t, \psi)\} \]

\[ \dot{U}_{t+1}(\psi) = U_{t+1}(\psi) - E\{U_{t+1}(\psi_0) \mid H_t\} \]

\[ Q_t = E\{U_{t+1}(\psi_0) \mid V_t\} - E\{U_{t+1}(\psi_0) \mid V_{t-1}\} \]

\[ S_t = \partial \log f(\sigma_{t+1} \mid V_t) / \partial \sigma_{t+1} \]

\[ W_t = \text{Var}(\sigma_{t+1} \mid V_t)^{-1}, \text{ which will be shown to be equal to } \text{Var}\{U_{t+1}(\psi_0) \mid V_t\}^{-1} \]

\[ T_t = E(W_t \mid H_t) \]

\[ T_t^* = E(T_t^{-1} \mid V_{t-1}) \]

\[ \epsilon_t = T_t^{-1}W_t\sigma_{t+1} + Q_t \]

\[ W_{t,t-1} = \text{Var}(\epsilon_t \mid V_{t-1})^{-1} \]

- \( \mathcal{H} \): the Hilbert space of all functions of \( V_{T+1} \) that have mean zero finite variance.

- \( \Lambda_1^t = \{ A_1^t = a_1^t(V_{T+1}) : E(A_1^t \mid V_t, Y_{t+1}) = 0 \} \)

- \( \Lambda_2^t = \{ A_2^t = a_2^t(\sigma_{t+1}, V_t) : E(A_2^t \mid V_t) = 0, E(A_2^t\sigma_{t+1} \mid V_t) = 0 \} \)

- \( \Lambda_3^t = \sum_{m=1}^t \Gamma_m^3 \)

- \( \Gamma_m^3 = \{ A_3^m = a_3^m(V_m) : E(A_3^m \mid H_m) = 0 \} \)

- \( \Lambda_4^t = \Gamma_4^t + \sum_{m=1}^{t-1} \Lambda_m^* \)

- \( \Lambda_m^* = \{ A_m^* = a_m^*(H_m) : E(A_m^* \mid V_{m-1}) = 0 \} \)

42
\[ \Gamma_t^4 = \{ A_t^* + S_t E(Q_t A_t^* | V_{t-1}) : A_t^* \in \Lambda_t^* \} \]

\[ \tilde{\Gamma}_t^4 = \{ A_t^* - E(Q_t A_t^* | V_{t-1})(T_t^*)^{-1}T_t^{-1}W_t \sigma_{t+1} : A_t^* \in \Lambda_t^* \} \]

\[ \Lambda_t^5 = \{ S_t A_t^* : A_t^* \in \Lambda_t^* \} \]

\[ \tilde{\Lambda}_t^5 = \{ A_t^* W_t \sigma_{t+1} : A_t^* \in \Lambda_t^* \} \]

\[ \Lambda_t^6 = \{ a(V_{t-1})S_t : a(V_{t-1}) \text{ is any function } \in \mathbb{R}^p \} \]

\[ \tilde{\Lambda}_t^6 = \{ a(V_{t-1}) \epsilon_t : a(V_{t-1}) \text{ is any function } \in \mathbb{R}^p \} \]

\[ D_t = E\{ h(\sigma_{t+1}, V_t) W_t \sigma_{t+1} | H_t \}, \text{ for a given } h(\sigma_{t+1}, V_t) \in \mathcal{H} \]

\[ R_t = E\{ B \sigma_{t+1} | V_t \} \text{ and } R_{t-1} = E\{ R_t W_t T_t^{-1} | V_{t-1} \}, \text{ for a given } B = b(V_{t+1}) \in \mathcal{H} \]

**F.4 Derivation of the general form of the efficient score**

**Lemma F.2.** Let \( \mathcal{M} \) denote the semiparametric model defined by consistency (Assumption [1]), positivity (Assumption [2]), (weak) sequential ignorability (Assumption [F.1]), and Assumption [F.2]. Let \( \mathcal{M}_t \) denote the semiparametric model defined by consistency, positivity, and the following \( t \)-specific version of (weak) sequential ignorability and [F.23]: for a fixed \( t \),

\[
\log \frac{E\{Y_{t+1}(\bar{a}_t) | \tilde{z}_t, \bar{a}_t\}}{E\{Y_{t+1}(\bar{a}_{t-1}, 0) | \tilde{z}_t, \bar{a}_t\}} = \gamma(t+1, \bar{z}_t, \bar{a}_t; \psi_0),
\]

\[ E\{Y_{t+1}(\bar{A}_{t-1}, 0) | H_t, A_t\} = E\{Y_{t+1}(\bar{A}_{t-1}, 0) | H_t\}. \]

Let \( \Lambda \) and \( \Lambda_t \) be the nuisance tangent space for model \( \mathcal{M} \) and model \( \mathcal{M}_t \), respectively. Then we have \( \mathcal{M} = \bigcap_{t=1}^T \mathcal{M}_t \) and \( \Lambda = \bigcap_{t=1}^T \Lambda_t \).

**Proof.** This follows directly from the definition of nuisance tangent space (i.e., \( L^2 \)-closure of all parametric submodel nuisance scores).
Lemma F.3. The nuisance tangent space for model $\mathcal{M}_t$ is $\Lambda_t = \Lambda_1^t + \Lambda_2^t + \Lambda_3^t + \Lambda_4^t + \Lambda_5^t + \Lambda_6^t$, where

\[
\begin{align*}
\Lambda_1^t &= \{ A_1^t = a_1^t(V_{T+1}) : E(A_1^t \mid V_t, Y_{t+1}) = 0 \}, \\
\Lambda_2^t &= \{ A_2^t = a_2^t(\sigma_{t+1}, V_t) : E(A_2^t \mid V_t) = 0, E(A_2^t \sigma_{t+1} \mid V_t) = 0 \}, \\
\Lambda_3^t &= \sum_{m=1}^{t} \Gamma_m^3, \\
\Lambda_4^t &= \Gamma_1^t + \sum_{m=1}^{t-1} \Lambda_m^t, \\
\Lambda_5^t &= \{ S_t A_t^t : A_t^t \in \Lambda_t^t \}, \\
\Lambda_6^t &= \{ a(V_{t-1}) S_t : a(V_{t-1}) \text{ is any function } \in \mathbb{R}^p \},
\end{align*}
\]

where

\[
\begin{align*}
\Gamma_m^3 &= \{ A_m^3 = a_m^3(V_m) : E(A_m^3 \mid H_m) = 0 \}, \\
\Lambda_m^t &= \{ A_m^t = a_m^t(H_m) : E(A_m^t \mid V_{m-1}) = 0 \} \\
\Gamma_1^t &= \{ A_1^t + S_t E(Q_t A_t^t \mid V_{t-1}) : A_t^t \in \Lambda_t^t \},
\end{align*}
\]

and

\[
\begin{align*}
Q_t &= E\{U_{t+1}(\psi_0) \mid V_t\} - E\{U_{t+1}(\psi_0) \mid V_{t-1}\}, \\
S_t &= \frac{\partial \log f(\sigma_{t+1} \mid V_t)}{\partial \sigma_{t+1}}.
\end{align*}
\]

Both $Q_t$ and $S_t$ are evaluated at the truth.

Proof. The likelihood for model $\mathcal{M}_t$ is

\[
L(\psi, \theta) = f(V_{T+1} \mid V_t, Y_{t+1}) f(Y_{t+1} \mid V_t) \prod_{m=1}^{t} \{ f(A_m \mid H_m) f(Z_m \mid V_{m-1}) \}
\]

\[
= f(V_{T+1} \mid V_t, Y_{t+1}; \theta_1) \\
\times \frac{\partial \sigma_{t+1}}{\partial Y_{t+1}} \times f(\sigma_{t+1}(\psi, \theta_4, \theta_5, \theta_6) \mid V_t; \theta_2) \\
\times \prod_{m=1}^{t} \{ f(A_m \mid H_m; \theta_3) f(Z_m \mid V_{m-1}; \theta_4) \},
\] (F.24)

44
where

\[ \sigma_{t+1}(\psi, \theta_4, \theta_5, \theta_6) = Y_{t+1} e^{-\gamma(t+1, H_t, A_t; \psi)} - \beta_t(V_{t-1}; \theta_6) - \left\{ q_t^*(H_t; \theta_5) - \int q_t^*(z_t, V_{t-1}; \theta_5) dF(z_t \mid V_{t-1}; \theta_4) \right\}. \]  \hspace{1cm} (F.25)

Here, \( \theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) \), each is an infinite-dimensional nuisance parameter and are variationally independent of each other. The second equality in (F.24) follows from the change of variables

\[ (L_1, A_1, Y_2, ..., Y_{t+1}, ..., L_T, A_T, Y_{T+1}) \rightarrow (L_1, A_1, Y_2, ..., \sigma_{t+1}, ..., L_T, A_T, Y_{T+1}) \]

which has Jacobian \( \partial \sigma_{t+1} / \partial Y_{t+1} = e^{-\gamma(t+1, H_t, A_t; \psi)} \). By Lemma G.2, the constraints on model \( \mathcal{M}_t \) is equivalent to \( E(\sigma_{t+1} \mid H_t, A_t) = 0 \), i.e., \( \int t dF(t \mid H_t, A_t) = 0 \). There is no restrictions on \( q_t^*(H_t; \theta_5) \) and \( \beta_t(H_{t-1}, A_{t-1}; \theta_6) \). The constraint \( E(q_t \mid H_{t-1}, A_{t-1}) = 0 \) has been incorporated because \( q_t^* \) is centered in (F.25).

Below we derive the nuisance tangent space for each nuisance parameter \( (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) \).

**Nuisance tangent space \( \Lambda_t^1 \) for \( \theta_1 \).** This follows from Theorem 4.6 in Tsiatis (2007).

**Nuisance tangent space \( \Lambda_t^2 \) for \( \theta_2 \).** This follows from Theorem 4.7 in Tsiatis (2007).

**Nuisance tangent space \( \Lambda_t^3 \) for \( \theta_3 \).** This follows from Theorem 4.6 in Tsiatis (2007).

**Nuisance tangent space \( \Lambda_t^4 \) for \( \theta_4 \).** The score for \( \theta_4 \) equals

\[ \frac{\partial \log L(\psi, \theta)}{\partial \theta_4} = \frac{\partial \log f(\sigma_{t+1}(\psi, \theta_4, \theta_5, \theta_6) \mid V_t; \theta_2)}{\partial \theta_4} + \sum_{m=1}^{t} \frac{\partial \log f(Z_m \mid V_{m-1}; \theta_4)}{\partial \theta_4}. \]

The \( \frac{\partial \log f(\sigma_{t+1}(\psi, \theta_4, \theta_5, \theta_6) \mid V_t; \theta_2)}{\partial \theta_4} + \frac{\partial \log f(Z_t \mid V_{t-1}; \theta_4)}{\partial \theta_4} \) part correspond to \( \Gamma_t^4 \), which is shown in the proof of Theorem A4.1 in Robins (1994). The \( \sum_{m=1}^{t-1} \frac{\partial \log f(Z_m \mid V_{m-1}; \theta_4)}{\partial \theta_4} \) part correspond to \( \sum_{m=1}^{t-1} \Lambda_t^4 \), which follows from Theorem 4.6 in Tsiatis (2007).

**Nuisance tangent space \( \Lambda_t^5 \) for \( \theta_5 \).** The score for \( \theta_5 \) equals

\[ \frac{\partial \log L(\psi, \theta)}{\partial \theta_5} = \frac{\partial \log f(\sigma_{t+1}(\psi, \theta_4, \theta_5, \theta_6) \mid V_t; \theta_2)}{\partial \theta_5}. \]

The form of \( \Lambda_t^5 \) is derived in the proof of Theorem A4.1 in Robins (1994).
Nuisance tangent space $\Lambda_t^6$ for $\theta_6$. The score for $\theta_6$ equals
\[
\frac{\partial \log L(\psi, \theta)}{\partial \theta_6} = \frac{\partial \log f(\sigma_{t+1}(\psi, \theta_4, \theta_5, \theta_6) | V_t; \theta_2)}{\partial \theta_6} = -\frac{\partial \log f(\sigma_{t+1}(\psi, \theta_4, \theta_5, \theta_6) | V_t; \theta_2)}{\partial \sigma_{t+1}} \times \frac{\partial \beta_t(V_{t-1}; \theta_6)}{\partial \theta_6}.
\]
Because there is no restriction on $\beta_t(V_{t-1}; \theta_6)$, $\Lambda_t^6 = \{a(V_{t-1})S_t: a(V_{t-1})$ is any function $\in \mathbb{R}^p\}$. \hfill \Box

Lemma F.4. The nuisance tangent space $\Lambda_t$ in Lemma G.3 equals the direct sum of the following spaces:
\[
\Lambda_t = \Lambda_t^1 \oplus \Lambda_t^2 \oplus \bigoplus_{m=1}^{t} \Gamma_m^3 \oplus \bigoplus_{m=1}^{t-1} \Lambda_m^4 \oplus \bigoplus_{m=1}^{t-1} \bar{\Lambda}_m^5 \oplus \bar{\Lambda}_t^6,
\]
where
\[
\bar{\Gamma}_t^4 = \{A_t^* - E(Q_t A_t^* | V_{t-1})(T_t^*)^{-1} T_t^{-1} W_t \sigma_{t+1} : A_t^* \in \Lambda_t^*, \}
\]
\[
\bar{\Lambda}_t^5 = \{A_t^* W_t \sigma_{t+1} : A_t^* \in \Lambda_t^*, \}
\]
\[
\bar{\Lambda}_t^6 = \{a(V_{t-1})\epsilon_t : a(V_{t-1}) \text{ is any function} \in \mathbb{R}^p\},
\]
and
\[
W_t = \text{Var}(\sigma_{t+1} | V_t)^{-1},
\]
\[
T_t = E(W_t | H_t),
\]
\[
T_t^* = E(T_t^{-1} | V_{t-1}),
\]
\[
\epsilon_t = T_t^{-1} W_t \sigma_{t+1} + Q_t.
\]

Proof. In Lemma G.3 we show that $\Lambda_t^1$, $\{\Gamma_m^3\}_{1 \leq m \leq t}$, $\{\Lambda_m^4\}_{1 \leq m \leq t-1}$ are orthogonal to each other and orthogonal to the rest subspaces $\Lambda_t^2, \bar{\Gamma}_t^4, \Lambda_t^5, \Lambda_t^6$. Thus, it suffices to show that $\bar{\Lambda}_t^5 = \Pi(\Lambda_t^5 | \Lambda_t^{2,4,5})$, $\bar{\Gamma}_t^4 = \Pi\{\Gamma_t^4 | (\Lambda_t^2 \oplus \bar{\Lambda}_t^5)\perp\}$, and $\bar{\Lambda}_t^6 = \Pi\{\Lambda_t^6 | (\Lambda_t^2 \oplus \bar{\Lambda}_t^5 \oplus \bar{\Gamma}_t^4)\perp\}$.

First, we show that $\bar{\Lambda}_t^5 = \Pi(\Lambda_t^5 | \Lambda_t^{2,4,5})$. For any $S_t A_t^* \in \Lambda_t^5$, because $S_t A_t^*$ is a function of $(\sigma_{t+1}, V_t)$, Lemma G.4 implies that
\[
\Pi(S_t A_t^* | \Lambda_t^5) = S_t A_t^* - E(S_t A_t^* \sigma_{t+1} | V_t) \text{Var}(\sigma_{t+1} | V_t)^{-1} \sigma_{t+1} - E(S_t A_t^* | V_t).
\]

(F.26)
By Lemma \ref{lem:G.4} we have $E(S_tA^*_t\sigma_{t+1} | V_t) = A^*_tE(S_t\sigma_{t+1} | V_t) = -A^*_t$ and $E(S_tA^*_t | V_t) = A^*_tE(S_t | V_t) = 0$, so \eqref{eq:F.26} implies

$$
\Pi(S_tA^*_t | \Lambda^2_t) = S_tA^*_t - \Pi(S_tA^*_t | \Lambda^2_t) = -A^*_tW_t\sigma_{t+1}.
$$

This gives the form of $\hat{A}^5_t$.

Second, we show that $\hat{\Gamma}^4_t = \Pi\{\Gamma^4_t | (\Lambda^2_t \oplus \hat{\Lambda}^5_t)^\perp\}$. For any $A^*_t + S_tE(Q_tA^*_t | V_{t-1}) \equiv g_1(\sigma_{t+1}, V_t) \in \Gamma^4_t$ where $A^*_t(H_t)$ satisfies $E(A^*_t | V_{t-1}) = 0$. By Lemma \ref{lem:G.12} it suffices to derive $\Pi\{\Pi(g_1 | \Lambda^2_t^\perp) | \hat{\Lambda}^5_t^\perp\}$. By Lemma \ref{lem:G.4} we have

$$
\Pi\{g_1(\sigma_{t+1}, V_t) | \Lambda^2_t\} = g_1 - E(g_1\sigma_{t+1} | V_t)\text{Var}(\sigma_{t+1} | V_t)^{-1}\sigma_{t+1} - E(g_1 | V_t).
$$

By Lemma \ref{lem:G.5} we have $E(g_1 | V_t) = A^*_t + E(S_t | V_t)E(Q_tA^*_t | V_{t-1}) = A^*_t$ and

$$
E(g_1\sigma_{t+1} | V_t) = A^*_tE(\sigma_{t+1} | V_t) + E(S_t\sigma_{t+1} | V_t)E(Q_tA^*_t | V_{t-1}) = -E(Q_tA^*_t | V_{t-1}).
$$

These combining with \eqref{eq:F.27} yields that

$$
\Pi\{g_1(\sigma_{t+1}, V_t) | \Lambda^2_t\} = g_1 - E(Q_tA^*_t | V_{t-1})W_t\sigma_{t+1}.
$$

Now, let $g_2(\sigma_{t+1}, V_t) = A^*_t - E(Q_tA^*_t | V_{t-1})W_t\sigma_{t+1}$. By Lemma \ref{lem:G.6} we have

$$
\Pi\{g_2(\sigma_{t+1}, V_t) | \hat{\Lambda}^5_t\} = \{-E(D_tT_t^{-1} | V_{t-1})(T_t^*T_t)^{-1} + D_tT_t^{-1}\}W_t\sigma_{t+1},
$$

where

$$
D_t = E\{g_2(\sigma_{t+1}, V_t) | W_t\sigma_{t+1} | H_t\}
$$

$$
= E(A^*_tW_t\sigma_{t+1} | H_t) - E\{E(Q_tA^*_t | V_{t-1})W^2_t\sigma^2_{t+1} | H_t\}
$$

$$
= -E(Q_tA^*_t | V_{t-1})T_t. \quad \text{(F.29)}
$$

The third equality in \eqref{eq:F.29} follows from Lemma \ref{lem:G.11} and the fact that $E(A^*_tW_t\sigma_{t+1} | H_t) = E\{A^*_tW_tE(\sigma_{t+1} | V_t) | H_t\} = 0$. Thus, plugging \eqref{eq:F.29} into \eqref{eq:F.28} and we have

$$
\Pi\{g_2(\sigma_{t+1}, V_t) | \hat{\Lambda}^5_t\} = \{E(Q_tA^*_t | V_{t-1})(T_t^*T_t)^{-1} - E(Q_tA^*_t | V_{t-1})\}W_t\sigma_{t+1}
$$

47
and
\[\Pi\{g_2(\sigma_{t+1}, V_t) | \tilde{\Lambda}^{5}_{t}\} = g_2(\sigma_{t+1}, V_t) - \Pi\{g_2(\sigma_{t+1}, V_t) | \tilde{\Lambda}^{5}_{t}\} = A_{t}^* - E(Q_t A_{t}^* | V_{t-1})(T_t^* T_t)^{-1} W_t \sigma_{t+1}.\]

This gives the form of \(\tilde{\Gamma}^4_t\).

Last, we show that \(\tilde{\Lambda}^0_t = \Pi\{\Lambda^0_t | (\Lambda^2_t \oplus \tilde{\Lambda}^5_t \oplus \tilde{\Gamma}^4_t)^{-1}\}.\) For any \(a(V_{t-1})S_t \in \Lambda^0_t\), by Lemma G.12 it suffices to derive \(\Pi(\Pi[\Pi\{a(V_{t-1})S_t | \Lambda^2_t \oplus \tilde{\Lambda}^5_t \oplus \tilde{\Gamma}^4_t\} | \tilde{\Lambda}^{5}_{t}\} | \tilde{\Gamma}^{4}_{t})\). By Lemma G.4 we have

\[\Pi\{a(V_{t-1})S_t | \Lambda^2_t\} = a(V_{t-1})S_t - E\{a(V_{t-1})S_t \sigma_{t+1} | V_t\} W_t \sigma_{t+1} - E\{a(V_{t-1})S_t | V_t\}\]

\[= a(V_{t-1})S_t + a(V_{t-1})W_t \sigma_{t+1},\]

where the second equality follows from Lemma G.5. Thus \(\Pi\{a(V_{t-1})S_t | \Lambda^2_t\} = -a(V_{t-1})W_t \sigma_{t+1} \equiv g_3(V_t, \sigma_{t+1}).\) By Lemma G.6 we have

\[\Pi\{g_3(\sigma_{t+1}, V_t) | \tilde{\Lambda}^{5}_{t}\} = \{-E(D_t^{(2)} T_t^{-1} | V_{t-1})(T_t^* T_t)^{-1} + D_t^{(2)} T_t^{-1}\} W_t \sigma_{t+1},\]

where, by using Lemma G.11 we have

\[D_t^{(2)} = E\{g_3(\sigma_{t+1}, V_t) W_t \sigma_{t+1} | H_t\}\]

\[= -a(V_{t-1})E(W_t^2 \sigma_{t+1}^2 | H_t)\]

\[= -a(V_{t-1})T_t.\]

Thus,

\[\Pi\{g_3(\sigma_{t+1}, V_t) | \tilde{\Lambda}^{5}_{t}\} = \{a(V_{t-1})(T_t^* T_t)^{-1} - a(V_{t-1})\} W_t \sigma_{t+1}\]

and

\[\Pi\{g_3(\sigma_{t+1}, V_t) | \tilde{\Lambda}^{5}_{t}\} = -a(V_{t-1})(T_t^* T_t)^{-1} W_t \sigma_{t+1} \equiv g_4(\sigma_{t+1}, V_t)\]

Note that \(E(g_4 | H_t) = E(g_4 | V_{t-1}) = 0\), so for \(O^*_3\) and \(O^*_4\) defined in Lemma G.6 we have

\[O^*_3 O^*_4(g_4) = -E\{g(T_t^* T_t^{-1} T_t^{-1} W_t \sigma_{t+1} | V_{t-1}) Q_t\}
\]

\[= E\{a(V_{t-1})(T_t^*)^{-2} T_t^{-2} W_t^2 \sigma_{t+1}^2 | V_{t-1}) Q_t\}
\]

\[= a(V_{t-1})(T_t^*)^{-2} E\{T_t^{-2} E(W_t^2 \sigma_{t+1}^2 | H_t) | V_{t-1}) Q_t\}
\]

\[= a(V_{t-1})(T_t^*)^{-2} E(T_t^{-1} | V_{t-1}) Q_t
\]

\[= a(V_{t-1})(T_t^*)^{-1} Q_t,\]

48
where the third to last equality follows from Lemma \([G.11]\) So by Lemma \([G.6]\) we have

\[
\Pi\{g_4(\sigma_{t+1}, V_t) \mid \tilde{\Gamma}_t^4\} = a(V_{t-1})(T_t^*)^{-1}Q_t - a(V_{t-1})(T_t^*)^{-1}E(Q_t^2 \mid V_{t-1})W_{t,t-1}\epsilon_t,
\]

where \(W_{t,t-1} = \text{Var}(\epsilon_t \mid V_{t-1})^{-1}\). Hence,

\[
\Pi\{g_4(\sigma_{t+1}, V_t) \mid \tilde{\Gamma}_t^4\} = g_4(\sigma_{t+1}, V_t) - \Pi\{g_4(\sigma_{t+1}, V_t) \mid \tilde{\Gamma}_t^4\} = -a(V_{t-1})(T_t^*)^{-1}(T_t^{-1}W_t\sigma_{t+1} + Q_t) + a(V_{t-1})(T_t^*)^{-1}E(Q_t^2 \mid V_{t-1})W_{t,t-1}\epsilon_t
\]

\[
= a(V_{t-1})(T_t^*)^{-1}\epsilon_t\{W_{t,t-1}E(Q_t^2 \mid V_{t-1}) - 1\}
\]

\[
= -a(V_{t-1})W_{t,t-1}\epsilon_t,
\]

where the last equality follows from Lemma \([G.11]\). Thus,

\[
\tilde{\Lambda}_t^6 = \Pi\{\Lambda_t^6 \mid (\Lambda_t^2 \oplus \tilde{\Lambda}_t^5 \oplus \tilde{\Gamma}_t^4)^{\perp}\} = \{-a(V_{t-1})W_{t,t-1}\epsilon_t : a(V_{t-1}) \text{ is any function } \in \mathbb{R}^p\} = \{a(V_{t-1})\epsilon_t : a(V_{t-1}) \text{ is any function } \in \mathbb{R}^p\},
\]

where the last equality follows from the fact that \(W_{t,t-1}\) is a function of \(V_{t-1}\).

This completes the proof. 

\(\square\)

**Lemma F.5.** For any \(B = b(V_{T+1}) \in \mathcal{H}\), its projection onto \(\Lambda_t^\perp\) is

\[
\Pi(B \mid \Lambda_t^\perp) = \{R_t - T_t^{-1}E(R_tW_t \mid H_t)\}W_t\sigma_{t+1},
\]

where \(R_t = E(B\sigma_{t+1} \mid V_t)\), and \(W_t, T_t\) are defined in Lemma \([F.4]\).

**Proof.** For any \(B = b(V_{T+1}) \in \mathcal{H}\), we have

\[
B = \{B - E(B \mid \sigma_{t+1}, V_t)\} + \{E(B \mid \sigma_{t+1}, V_t) - E(B \mid V_t)\}
\]

\[
+ \sum_{m=1}^{t}\{E(B \mid V_m) - E(B \mid H_m)\} + \sum_{m=1}^{t}\{E(B \mid H_m) - E(B \mid V_{m-1})\}.
\]

Note that \(B - E(B \mid \sigma_{t+1}, V_t) \in \Lambda_t^1\), and for all, \(1 \leq m \leq t\) \(E(B \mid V_m) - E(B \mid H_m) \in \Gamma_m^3\) and \(E(B \mid H_m) - E(B \mid V_{m-1}) \in \Lambda_m^*\). Hence, by Lemma \([F.4]\) we have

\[
\Pi(B \mid \Lambda_t^\perp) = \Pi\{E(B \mid \sigma_{t+1}, V_t) - E(B \mid V_t) \mid \Lambda_t^\perp\} + \Pi\{E(B \mid H_t) - E(B \mid V_{t-1}) \mid \Lambda_t^\perp\}. \quad (F.30)
\]
By Lemma \[\text{G.9}\] we have
\[
\Pi\{E(B \mid \sigma_{t+1}, V_t) - E(B \mid V_t) \mid \Lambda_t^\perp\} = \{R_t - T_t^{-1}E(R_t W_t \mid H_t)\} W_t \sigma_{t+1},
\]
where \(R_t = E(B\sigma_{t+1} \mid V_t)\). By Lemma \[\text{G.10}\], we have \(\Pi\{E(B \mid H_t) - E(B \mid V_{t-1}) \mid \Lambda_t^\perp\} = 0\). Plugging those into \(\text{F.30}\) completes the proof.

\[\text{Lemma F.6.}\] The orthogonal complement of the nuisance tangent space, \(\Lambda_t^\perp\), is
\[
\Lambda_t^\perp = \{d(V_t)\sigma_{t+1} : \text{any } d(V_t) \in \mathbb{R}^p \text{ such that } E[d(V_t) \mid H_t] = 0\}.
\]

\[\text{Proof.}\] Lemma \[\text{F.5}\] implies that
\[
\Lambda_t^\perp = \{[R_t - T_t^{-1}E(R_t W_t \mid H_t)] W_t \sigma_{t+1} : R_t = E(h\sigma_{t+1} \mid V_t), h \in \mathcal{H}\}.
\]
Denote by \(\Lambda_t^{\perp, \text{conj}} = \{d(V_t)\sigma_{t+1} : \text{any } d(V_t) \text{ such that } E[d(V_t) \mid H_t] = 0\}\). In the following we show \(\Lambda_t^\perp = \Lambda_t^{\perp, \text{conj}}\).

First we show \(\Lambda_t^\perp \subset \Lambda_t^{\perp, \text{conj}}\). For any \(h \in \mathcal{H}\), we have
\[
E[\{R_t - T_t^{-1}E(R_t W_t \mid X_t)\} W_t \sigma_{t+1} \mid H_t] = 0,
\]
and hence \(\Lambda_t^\perp \subset \Lambda_t^{\perp, \text{conj}}\).

Next we show \(\Lambda_t^{\perp, \text{conj}} \subset \Lambda_t^\perp\). For any \(d(V_t)\sigma_{t+1} \in \Lambda_t^{\perp, \text{conj}}\), i.e. for any \(d(V_t)\) such that \(E[d(V_t) \mid H_t] = 0\), let \(h = d(V_t)\sigma_{t+1} \in \mathcal{H}\), and we have
\[
R_t W_t = E(h\sigma_{t+1} \mid V_t) W_t = d(V_t) E(\sigma_{t+1}^2 \mid V_t) W_t = d(V_t),
\]
and so \(E(R_t W_t \mid H_t) = 0\). Therefore,
\[
\{R_t - T_t^{-1}E(R_t W_t \mid H_t)\} W_t \sigma_{t+1} = R_t W_t \sigma_{t+1} = d(V_t)\sigma_{t+1}.
\]
This implies that \(d(V_t)\sigma_{t+1} \in \Lambda_t^\perp\), and hence \(\Lambda_t^{\perp, \text{conj}} \subset \Lambda_t^\perp\). This completes the proof.

\[\text{Lemma F.7.}\] The orthogonal complement of the nuisance tangent space for model \(\mathcal{M}\) defined in Lemma \[\text{F.2}\] is
\[
\Lambda_t^\perp = \{\sum_{t=1}^T d_t(V_t)\dot{U}_{t+1}(\psi_0) : \text{any } d_t(V_t) \in \mathbb{R}^p \text{ s.t. } E[d_t(V_t) \mid H_t] = 0\}, \quad (\text{F.31})
\]
where \(U_{t+1}(\psi) = Y_{t+1}\exp\{-\gamma(t+1, V_t; \psi_0)\}\) and \(\dot{U}_{t+1}(\psi) = U_{t+1}(\psi) - E[U_{t+1}(\psi_0) \mid H_t]\).
Proof. Lemma [F.6] implies that

\[ \Lambda_t^+ = \{ d(V_t) | \{ U_{t+1}(\psi) \} \} : \text{any } d(V_t) \in \mathbb{R}^p \text{ such that } E[d(V_t) | H_t] = 0. \]

Because \((\cap_{t=1}^T \Lambda_t^+) = \sum_{t=1}^T \Lambda_t^+ \), ([F.31]) is an immediate implication of Lemma [F.2].

**Lemma F.8.** The efficient score \( S_{\text{eff}}(\psi) \) is

\[
S_{\text{eff}}(\psi) = - \sum_{t=1}^T \left[ E \left\{ \frac{\partial U_{t+1}(\psi)}{\partial \psi} | V_t \right\} - E \left\{ \frac{\partial U_{t+1}(\psi)}{\partial \psi} W_t | H_t \right\} E(W_t | H_t)^{-1} \right] W_t \dot{U}_{t+1}(\psi),
\]

where \( W_t = \text{Var} \{ U_{t+1}(\psi) | V_t \}^{-1} \).

**Proof.** By definition, the efficient score is the projection of the score for \( \psi, S_\psi \), onto \( \Lambda^\perp \). For any \( 1 \leq t < s \leq T \), for any \( d_t(V_t) \dot{U}_{t+1}(\psi) \in \Lambda_t^+ \) and \( d_s(V_s) \dot{U}_{s+1}(\psi) \in \Lambda_s^+ \), their inner product is

\[
E \{ d_t(V_t) \dot{U}_{t+1}(\psi) d_s(V_s) \dot{U}_{s+1}(\psi) \} = E \{ d_t(V_t) \dot{U}_{t+1}(\psi) d_s(V_s) \} E \{ \dot{U}_{s+1}(\psi) | V_s \} = 0,
\]

where the last equality follows from Lemma [G.1]. This implies that \( \Lambda_t^+ \perp \Lambda_s^+ \) for any \( 1 \leq t < s \leq T \). Therefore, \( \Lambda^\perp = \bigoplus_{t=1}^T \Lambda_t^\perp \), and \( \Pi(S_\psi | \Lambda^\perp) = \sum_{t=1}^T \Pi(S_\psi | \Lambda_t^\perp) \). By Lemma [F.5], we have

\[
\Pi(S_\psi | \Lambda_t^\perp) = \{ R_t - T_t^{-1} E(R_t W_t | H_t) \} W_t \sigma_{t+1}
\]

\[
= \{ E(S_\psi \sigma_{t+1} | V_t) - E(S_\psi \sigma_{t+1} W_t | H_t) E(W_t | H_t)^{-1} \} W_t \sigma_{t+1}.
\] (F.32)

We have \( \sigma_{t+1} = U_{t+1}(\psi) - E_P \{ U_{t+1}(\psi) | V_t \} = \bar{U}_{t+1}(\psi) \) as in Lemma [G.2], so \( W_t = \text{Var}(\sigma_{t+1} | V_t)^{-1} = \text{Var} \{ U_{t+1}(\psi) | V_t \}^{-1} \). By the generalized information equality ([Newey 1990])

\[
E(S_\psi \dot{U}_{t+1}(\psi) | V_t) = - E \left\{ \frac{\partial \bar{U}_{t+1}(\psi)}{\partial \psi} | V_t \right\} = - E \left\{ \frac{\partial U_{t+1}(\psi)}{\partial \psi} | V_t \right\},
\]

So ([F.32]) becomes

\[
\Pi(S_\psi | \Lambda_t^\perp) = \{ E(S_\psi \sigma_{t+1} | V_t) - E(S_\psi \sigma_{t+1} W_t | H_t) E(W_t | H_t)^{-1} \} W_t \sigma_{t+1}
\]

\[
= - \left[ E \left\{ \frac{\partial U_{t+1}(\psi)}{\partial \psi} | V_t \right\} - E \left\{ \frac{\partial U_{t+1}(\psi)}{\partial \psi} W_t | H_t \right\} E(W_t | H_t)^{-1} \right] W_t \dot{U}_{t+1}(\psi).
\]

Thus, the form of \( S_{\text{eff}}(\psi) \) follows from the fact that \( S_{\text{eff}}(\psi) = \Pi(S_\psi | \Lambda^\perp) = \sum_{t=1}^T \Pi(S_\psi | \Lambda_t^\perp) \).
**G  Supporting lemmas used in Section F**

**Lemma G.1.** \( E\{U_{t+1}(\psi_0) \mid H_t, A_t\} = E\{U_{t+1}(\psi_0) \mid H_t\} \).

*Proof.* We have

\[
E\{H_{t+1}(\psi) \mid H_t, A_t\} = E\{Y_{t+1} \mid H_t, A_t\}
\]

(consistency) \( = E\{Y_{t+1}(A_t) \mid H_t, A_t\}e^{-\gamma(t+1,Z_t,A_t;\psi)} \)

(by (F.23)) \( = E\{Y_{t+1}(A_{t-1},0) \mid H_t, A_t\}e^{\gamma(t+1,Z_t,A_t;\psi_0) - \gamma(t+1,Z_t,A_t;\psi)} \)

(sequential ignorability) \( = E\{Y_{t+1}(A_{t-1},0) \mid H_t\}e^{\gamma(t+1,Z_t,A_t;\psi_0) - \gamma(t+1,Z_t,A_t;\psi)} \).

Therefore \( E\{U_{t+1}(\psi_0) \mid H_t, A_t\} = E\{U_{t+1}(\psi_0) \mid H_t\}. \) \( \square \)

**Lemma G.2.** Let \( \sigma_{t+1} \) be a random variable that is defined on the same sample space as \( V_{T+1} \). Consider a tuple \((P', q_t(H_t), \beta_t(H_{t-1}, A_{t-1}))\), where \( P' \) is a probability distribution of \( V_{T+1} \cup \sigma_{t+1} \setminus Y_{t+1} \), \( q_t \) is a (deterministic) function of \( H_t \), and \( \beta_t \) is a (deterministic) function of \( H_{t-1}, A_{t-1} \). Define \( \mathcal{M}'_t \) the collection of \((P', q_t(H_t), \beta_t(H_{t-1}, A_{t-1}))\) tuples such that positivity holds for \( P' \) and that

\[
E(\sigma_{t+1} \mid H_t, A_t) = 0, \tag{G.1}
\]

\[
q_t(H_t, A_t) = q_t(H_t) \text{ is a function of } H_t, \tag{G.2}
\]

\[
E(q_t \mid H_{t-1}, A_{t-1}) = 0. \tag{G.3}
\]

Then there is a 1-1 mapping \( g \) between \( \mathcal{M}_t \) and \( \mathcal{M}'_t \) given by:

\[
g : P \mapsto (P', q_t(H_t), \beta_t(H_{t-1}, A_{t-1}))
\]

where \( P' \) is induced by \( P \) and \( \sigma_{t+1} = U_{t+1}(\psi_0) - E_P\{U_{t+1}(\psi_0) \mid H_t, A_t\}, q_t(H_t, A_t) = E_P\{U_{t+1}(\psi_0) \mid H_t, A_t\} - E_P\{U_{t+1}(\psi_0) \mid H_{t-1}, A_{t-1}\} \) and \( \beta_t(H_{t-1}, A_{t-1}) = E_P\{U_{t+1}(\psi_0) \mid H_{t-1}, A_{t-1}\} \). The inverse mapping is

\[
g^{-1} : (P', q_t(H_t), \beta_t(H_{t-1}, A_{t-1})) \mapsto P
\]

where \( P \) is induced by \( P' \) and \( Y_{t+1} = e^{\gamma(t+1,H_t,A_t;\psi_0)}(\sigma_{t+1} + q_t + \beta_t) \).

52
Proof. First, we show that the $g(P) \in \mathcal{M}_t'$. Let $q_t = E\{U_{t+1}(\psi_0) \mid H_t, A_t\} - E\{U_{t+1}(\psi_0) \mid H_{t-1}, A_{t-1}\}$, $\beta_t = E\{U_{t+1}(\psi_0) \mid H_{t-1}, A_{t-1}\}$, and $\sigma_{t+1} = Y_{t+1} - q_t - \beta_t$. Let $P'$ be the probability distribution of $V_{T+1} \cup \sigma_{t+1} \setminus Y_{t+1}$ induced by $\sigma_{t+1} = Y_{t+1} e^{-\gamma(t+1, H_t, A_t; \psi_0)} - q_t - \beta_t$ and $P$. Trivially we have $E(\sigma_{t+1} \mid H_t, A_t) = 0$ and $E(q_t \mid H_{t-1}, A_{t-1}) = 0$. Because $P \in \mathcal{M}_t$, Lemma G.1 implies $q_t = q_t(H_t)$. Therefore, $(P', q_t, \beta_t) \in \mathcal{M}_t'$.

Then we show that $g^{-1}\{P', q_t(H_t), \beta_t(H_{t-1}, A_{t-1})\} \in \mathcal{M}_t$. We have

$$E\{Y_{t+1} e^{-\gamma(t+1, H_t, A_t; \psi_0)} \mid H_t, A_t\}$$

$$= E\{\sigma_{t+1} + q_t + \beta_t \mid H_t, A_t\}$$

$$= 0 + q_t(H_t) + \beta_t(H_{t-1}, A_{t-1}).$$

Set $A_t = 0$, we have weak sequential ignorability

$$E\{Y_{t+1}(\tilde{A}_{t-1}, 0) \mid H_t, A_t = 0\} = E\{Y_{t+1}(\tilde{A}_{t-1}, 0) \mid H_t\}.$$

Taking the ratio between $A_t = a_t$ and $A_t = 0$, we have (F.23)

$$\frac{E\{Y_{t+1}(\tilde{A}_{t-1}, a_t) \mid H_t, A_t\}}{E\{Y_{t+1}(\tilde{A}_{t-1}, 0) \mid H_t, A_t\}} = e^{\gamma(t+1, H_t, a_t; \psi_0)}.$$

Therefore $P \in \mathcal{M}_t$. $\Box$

Lemma G.3. $\Lambda_t^1$, $\{\Gamma_m^3\}_{1 \leq m \leq t}$, $\{\Lambda_m^*\}_{1 \leq m \leq t-1}$ are orthogonal to each other and orthogonal to the rest subspaces $\Lambda_t^2, \Gamma_t^4, \Lambda_t^5, \Lambda_t^6$.

Proof. Using the definition in Lemma F.3, we have the following. (We will repeatedly use the fact $E(S_t \mid V_t) = 0$, which is shown in Lemma G.5)

- $\Lambda_t^1 \perp \Lambda_t^2$: $\forall A_t^1 \in \Lambda_t^1, A_t^2 \in \Lambda_t^2$, we have

$$E(A_t^1 A_t^2) = E\{E(A_t^1 A_t^2 \mid V_t, Y_{t+1})\} = E\{A_t^2 E(A_t^1 \mid V_t, Y_{t+1})\} = 0.$$

Similarly, we can show $\Lambda_t^1 \perp \Gamma_m^3, \Lambda_t^1 \perp \Gamma_t^4, \Lambda_t^1 \perp \Lambda_m^*, \Lambda_t^1 \perp \Lambda_t^5, \Lambda_t^1 \perp \Lambda_t^6$.

- $\Gamma_m^3 \perp \Lambda_t^2$ for all $1 \leq m \leq t$: $\forall A_m^3 \in \Gamma_m^3, A_t^2 \in \Lambda_t^2$, we have

$$E(A_m^3 A_t^2) = E\{E(A_m^3 A_t^2 \mid V_t)\} = E\{A_m^3 E(A_t^2 \mid V_t)\} = 0.$$
\(\Gamma_m^3 \perp \Gamma_k^3\) for all \(1 \leq m < k \leq t\): \(\forall A_m^3 \in \Gamma_m^3, A_k^3 \in \Gamma_k^3\), we have

\[
E(A_m^3 A_k^3) = E\{E(A_m^3 A_k^3 | H_k)\} = E\{A_m^3 E(A_k^3 | H_k)\} = 0.
\]

\(\Gamma_m^3 \perp \Gamma_t^4\) for all \(1 \leq m \leq t - 1\): \(\forall A_m^3 \in \Gamma_m^3, A_t^* + S_t E(Q_t A_t^* | V_{t-1}) \in \Gamma_t^4\), we have

\[
E[A_m^3\{A_t^* + S_t E(Q_t A_t^* | V_{t-1})\}] = E(A_m^3 A_t^*) + E\{A_m^3 S_t E(Q_t A_t^* | V_{t-1})\}
\]

\[
= E\{E(A_m^3 A_t^* | V_{t-1})\} + E[E\{A_m^3 S_t E(Q_t A_t^* | V_{t-1}) | V_t\}]
\]

\[
= E\{A_m^3 E(A_t^* | V_{t-1})\} + E[A_m^3 E(Q_t A_t^* | V_{t-1}) E(S_t | V_t)]
= 0
\]

\(\Gamma_t^3 \perp \Gamma_t^4\): \(\forall A_t^3 \in \Gamma_t^3, A_t^* + S_t E(Q_t A_t^* | V_{t-1}) \in \Gamma_t^4\), we have

\[
E[A_t^3\{A_t^* + S_t E(Q_t A_t^* | V_{t-1})\}] = E(A_t^3 A_t^*) + E\{A_t^3 S_t E(Q_t A_t^* | V_{t-1})\}
\]

\[
= E\{E(A_t^3 A_t^* | H_t)\} + E[E\{A_t^3 S_t E(Q_t A_t^* | V_{t-1}) | V_t\}]
\]

\[
= E\{A_t^* E(A_t^3 | H_t)\} + E[A_t^3 E(Q_t A_t^* | V_{t-1}) E(S_t | V_t)]
= 0
\]

\(\Gamma_m^3 \perp \Lambda_k^3\) for all \(1 \leq m \leq t\) and \(1 \leq k \leq t - 1\): \(\forall A_m^3 \in \Gamma_m^3, A_k^3 \in \Lambda_k^3\), if \(m < k\) we have

\[
E(A_m^3 A_k^3) = E\{E(A_m^3 A_k^3 | V_{k-1})\} = E\{A_m^3 E(A_k^3 | V_{k-1})\} = 0;
\]

if \(m \geq k\) we have

\[
E(A_m^3 A_k^3) = E\{E(A_m^3 A_k^3 | H_m)\} = E\{A_k^3 E(A_m^3 | H_m)\} = 0.
\]

\(\Gamma_m^3 \perp \Lambda_t^5\) for all \(1 \leq m \leq t\): \(\forall A_m^3 \in \Gamma_m^3, S_t A_t^* \in \Lambda_t^5\), we have

\[
E(A_m^3 S_t A_t^*) = E\{E(A_m^3 S_t A_t^* | V_t)\} = E\{A_m^3 A_t^* E(S_t | V_t)\} = 0.
\]

\(\Gamma_m^3 \perp \Lambda_t^6\) for all \(1 \leq m \leq t\): \(\forall A_m^3 \in \Gamma_m^3, a(V_{t-1}) S_t \in \Lambda_t^6\), we have

\[
E\{A_m^3 a(V_{t-1}) S_t\} = E[E\{A_m^3 a(V_{t-1}) S_t | V_t\}] = E\{A_m^3 a(V_{t-1}) E(S_t | V_t)\} = 0.
\]

54
• $\Lambda_m \perp \Lambda_t^2$ for all $1 \leq m \leq t-1$: $\forall A_m^* \in \Lambda_m^*, A_t^2 \in \Lambda_t^2$, we have

$$E(A_m^* A_t^2) = E\{E(A_m^* A_t^2 \mid V_t)\} = E\{A_m^* E(A_t^2 \mid V_t)\} = 0.$$

• $\Lambda_m^* \perp \Gamma_t^t$ for all $1 \leq m \leq t-1$: $\forall A_m^* \in \Lambda_m^*, A_t^* \in \Gamma_t^t$, we have

$$E[A_m^* \{A_t^* + S_t E(Q_t A_t^* \mid V_{t-1})\}] = E(A_m^* A_t^*) + E\{A_m^* S_t E(Q_t A_t^* \mid V_{t-1}) \mid V_t\}
\quad = E\{E(A_m^* \mid V_{t-1})\} + E\{E(A_m^* S_t E(Q_t A_t^* \mid V_{t-1}) \mid V_t)\}
\quad = E\{A_m^* E(A_t^* \mid V_{t-1})\} + E\{A_m^* E(Q_t A_t^* \mid V_{t-1}) E(S_t \mid V_t)\}
\quad = 0.$$

• $\Lambda_m^* \perp \Lambda_k^*$ for all $1 \leq m < k \leq t-1$: $\forall A_m^* \in \Lambda_m^*, A_k^* \in \Lambda_k^*$, we have

$$E(A_m^* A_k^*) = E\{E(A_m^* A_k^* \mid V_{k-1})\} = E\{A_m^* E(A_k^* \mid V_{k-1})\} = 0.$$

• $\Lambda_m^* \perp \Lambda_t^\circ$ for all $1 \leq m \leq t-1$: $\forall A_m^* \in \Lambda_m^*, S_t A_t^* \in \Lambda_t^\circ$, we have

$$E(A_m^* S_t A_t^*) = E\{E(A_m^* S_t A_t^* \mid V_t)\} = E\{A_m^* A_t^* E(S_t \mid V_t)\} = 0.$$

• $\Lambda_m^* \perp \Lambda_t^\circ$ for all $1 \leq m \leq t-1$: $\forall A_m^* \in \Lambda_m^*, a(V_{t-1}) S_t \in \Lambda_t^\circ$, we have

$$E\{A_m^* a(V_{t-1}) S_t\} = E[E\{A_m^* a(V_{t-1}) S_t \mid V_t\}] = E\{A_m^* a(V_{t-1}) E(S_t \mid V_t)\} = 0.$$

\[\square\]

**Lemma G.4** (Projection onto $\tilde{\Lambda}_T^2$). Let $\mathcal{G}$ be the Hilbert space of all mean-zero finite-variance functions of $(X,Y)$, where $(X,Y)$ follows some unknown distribution $P$ and the only restriction is $E(X \mid Y) = 0$. Let

$$\Lambda = \{h(X,Y) \in \mathcal{G} : E(h \mid Y) = 0, E(h X \mid Y) = 0\},$$

then we have $\Lambda = \{O(h) : h \in \mathcal{G}\} = O(\mathcal{G})$, where the operator $O = O_2 \circ O_1$ with

$$O_1(h) = h - E(h \mid Y),$$

$$O_2(h) = h - E(h X \mid Y) \text{Var}(X \mid Y)^{-1} X.$$
Both $O_1$ and $O_2$ are self-adjoint, i.e., $O_1^*=O_1$ and $O_2^*=O_2$. For any $h(X,Y) \in \mathcal{G}$, its projection onto $\Lambda$ equals

$$\Pi\{h(X,Y) \mid \Lambda\} = h - E(hX \mid Y)\text{Var}(X \mid Y)^{-1}X - E(h \mid Y).$$

**Proof.** The first statement in the Lemma is that $\Lambda = \Lambda' = O(\mathcal{G})$. To show this, we first show $\Lambda' \subset \Lambda$. For any $h \in \mathcal{G}$, i.e., for any $O(h) \in \Lambda'$, we have

$$E\{O(h) \mid Y\} = E[O_1(h) - E\{O_1(h)X \mid Y\}\text{Var}(X \mid Y)^{-1}X \mid Y]$$

$$= E\{O_1(h) \mid Y\} - E\{O_1(h)X \mid Y\}\text{Var}(X \mid Y)^{-1}E(X \mid Y)$$

$$= 0 - 0 = 0,$$

and

$$E\{O(h)X \mid Y\} = E[O_1(h)X - E\{O_1(h)X \mid Y\}\text{Var}(X \mid Y)^{-1}X^2 \mid Y]$$

$$= E\{O_1(h)X \mid Y\} - E\{O_1(h)X \mid Y\} = 0,$$

so $O(h) \in \Lambda$ and $\Lambda' \subset \Lambda$. Next, we show $\Lambda \subset \Lambda'$, i.e., for any $h \in \Lambda$, there exists $g \in \mathcal{G}$ such that $O(g) = h$. We claim that

$$O(g) = g - E(g \mid Y) - E\{[g - E(g \mid Y)]X \mid Y\}\text{Var}(X \mid Y)^{-1}X$$

$$= g - E(g \mid Y) - E(gX \mid Y)\text{Var}(X \mid Y)^{-1}X + E(g \mid Y)E(X \mid Y)\text{Var}(X \mid Y)^{-1}X \quad \text{(G.4)}$$

Because $h \in \Lambda$, we have $E(h \mid Y) = 0$ and $E(hX \mid Y) = 0$. Therefore, let $g = h$ in (G.4) and it becomes $O(h) = h$, and thus $\Lambda \subset \Lambda'$.

Next we show that $O_1$ and $O_2$ are both self-adjoint. For any $h,g \in \mathcal{G}$, we have

$$<O_1(h),g> = E\{[h - E(h \mid Y)]g\} = E(hg) - E\{E(h \mid Y)g\}$$

$$= E(hg) - E\{hE(g \mid Y)\} = E[h\{g - E(g \mid Y)\}],$$

56
and

\[ <O_2(h), g> = E\{h - E(hX|Y)\text{Var}(X|Y)^{-1}X\}g \]
\[ = E(hg) - E\{E(hX|Y)\text{Var}(X|Y)^{-1}Xg|Y]\}
\[ = E(hg) - E\{E(hX|Y)\text{Var}(X|Y)^{-1}E(Xg|Y)\}
\[ = E(hg) - E\{E(hX|Y)^{-1}E(Xg|Y)|Y\}
\[ = E[h\{g - X\text{Var}(X|Y)^{-1}E(Xg|Y)\}].\]

Hence both \(O_1\) and \(O_2\) are both self-adjoint.

The adjoint operator for \(O\) is

\[ O^*(h) = O_1^* \circ O_2^*(h) \]
\[ = h - E(hX|Y)\text{Var}(X|Y)^{-1}X - E\{h - E(hX|Y)\text{Var}(X|Y)^{-1}X|Y\}
\[ = h - E(hX|Y)\text{Var}(X|Y)^{-1}X - E(h|Y).\]

By a functional analysis result, for any \(h \in \mathcal{G}\), its projection \(\Pi(h|\Lambda)\) satisfies \(O^*\{\Pi(h|\Lambda)\} = O^*(h)\), i.e.,

\[ \Pi(h|\Lambda) - E\{\Pi(h|\Lambda)X|Y\}\text{Var}(X|Y)^{-1}X - E\{\Pi(h|\Lambda)|Y\}
\[ = h - E(hX|Y)\text{Var}(X|Y)^{-1}X - E(h|Y). \tag{G.5} \]

Because \(\Pi(h|\Lambda) \in \Lambda\), \(E\{\Pi(h|\Lambda)X|Y\} = E\{\Pi(h|\Lambda)|Y\} = 0\), so (G.5) yields

\[ \Pi(h|\Lambda) = h - E(hX|Y)\text{Var}(X|Y)^{-1}X - E(h|Y). \]

This completes the proof. \(\square\)

**Lemma G.5.** Consider a random variable \(X\) with \(E(X) = 0\). Let \(S(x) = \partial \log p(x)/\partial x\) where \(p(x)\) is the density of \(X\). Then under regularity conditions, \(E\{S(X)\} = 0\) and \(E\{S(X)X\} = -1\).

**Proof.** We have

\[ E\{S(X)\} = \int_{-\infty}^{\infty} p(x) \frac{\partial \log p(x)}{\partial x} dx = \int_{-\infty}^{\infty} \frac{\partial p(x)}{\partial x} dx \]
\[ = \int_{-\infty}^{\infty} \frac{\partial p(x+u)}{\partial u} \bigg|_{u=0} dx = \frac{\partial}{\partial u} \left\{ \int_{-\infty}^{\infty} p(x+u) dx \right\} \bigg|_{u=0} = 0 \]
F.4. For any mean zero function $g$, any $\tilde{\Lambda}$ as defined in Lemma F.4.

Proof. By the definition of $T$, we have

$$E\{S(X)X\} = \int_{-\infty}^{\infty} p(x) \frac{\partial \log p(x)}{\partial x} dx = \int_{-\infty}^{\infty} \frac{\partial p(x)}{\partial x} dx$$

$$= \left[ \int_{-\infty}^{\infty} \frac{\partial p(x+u)}{\partial u} x \right]_{u=0} dx = \left. \frac{\partial}{\partial u} \left\{ \int_{-\infty}^{\infty} p(x+u) dx \right\} \right|_{u=0}$$

$$= \left. \frac{\partial}{\partial u} \left\{ \int_{-\infty}^{\infty} p(t) (t-u) dt \right\} \right|_{u=0} = \left. \frac{\partial}{\partial u} (-u) \right|_{u=0} = -1.$$ 

\[ \square \]

Lemma G.6 (Projection onto $\tilde{\Lambda}_t^5$). Consider $\tilde{\Lambda}_t^5 = \{ \tilde{A}_t^* W_t \sigma_{t+1} : \tilde{A}_t^* \in \Lambda_t^* \}$ as defined in Lemma F.4. For any mean zero function $h(\sigma_{t+1}, V_t) \in \mathcal{H}$, its projection onto $\tilde{\Lambda}_t^5$ is

$$\Pi \{ h(\sigma_{t+1}, V_t) \mid \tilde{\Lambda}_t^5 \} = \tilde{A}_t^* W_t \sigma_{t+1},$$

where $\tilde{A}_t^* = -E(D_t T_{t}^{-1} | V_{t-1}) (T_{t}^{*} T_{t})^{-1} + D_t T_{t}^{-1}$, with $D_t = E\{ h(\sigma_{t+1}, V_t) W_t \sigma_{t+1} \mid H_t \}$ and $W_t, T_t, T_{t}^{*}$ as defined in Lemma F.4.

Proof. We can express $\tilde{\Lambda}_t^5$ as the image of the operator $O_2 \circ O_1$: $\tilde{\Lambda}_t^5 = O_2 \circ O_1(\mathcal{H})$, where for any $g(V_{t+1}) \in \mathcal{H}$ define $O_1(g) = E(g \mid H_t) - E(g \mid V_{t-1})$ and $O_2(g) = g W_t \sigma_{t+1}$. It follows that both $O_1$ and $O_2$ are self-adjoint. Suppose $\Pi \{ h(\sigma_{t+1}, V_t) \mid \tilde{\Lambda}_t^5 \} = \tilde{A}_t^* W_t \sigma_{t+1}$ for some $\tilde{A}_t^* (H_t) \in \Lambda_t^*$. By a functional analysis result, we have $O_1^* \circ O_2^* (\tilde{A}_t^* W_t \sigma_{t+1}) = O_1^* \circ O_2^* \{ h(\sigma_{t+1}, V_t) \}$, i.e.,

$$E(\tilde{A}_t^* W_t^2 \sigma_{t+1}^2 \mid H_t) = E(h W_t \sigma_{t+1} \mid H_t) = E(h W_t \sigma_{t+1} \mid V_{t-1}). \tag{G.6}$$

By Lemma G.11, $E(\tilde{A}_t^* W_t^2 \sigma_{t+1} \mid H_t) = \tilde{A}_t^* T_t$ and $E(\tilde{A}_t^* W_t^2 \sigma_{t+1}^2 \mid V_{t-1}) = E(\{E(\tilde{A}_t^* W_t^2 \sigma_{t+1} \mid H_t)\mid V_{t-1}\}) = E(\tilde{A}_t^* T_t \mid V_{t-1})$. So by the definition of $D_t$, (G.6) becomes $\tilde{A}_t^* T_t = E(\tilde{A}_t^* T_t \mid V_{t-1}) = D_t - E(D_t \mid V_{t-1})$, i.e.,

$$\tilde{A}_t^* = T_t^{-1} \{ E(\tilde{A}_t^* T_t \mid V_{t-1}) + D_t - E(D_t \mid V_{t-1}) \}. \tag{G.7}$$

Because $\tilde{A}_t^* \in \Lambda_t^*$, $E(\tilde{A}_t^* \mid V_{t-1}) = 0$, so taking $E(\cdot \mid V_{t-1})$ on both sides of (G.7) gives

$$0 = E(T_t^{-1} \mid V_{t-1}) E(\tilde{A}_t^* T_t \mid V_{t-1}) + E(T_t^{-1} D_t \mid V_{t-1}) - E(T_t^{-1} \mid V_{t-1}) E(D_t \mid V_{t-1}).$$

By the definition of $T_t^*$ this becomes

$$E(\tilde{A}_t^* T_t \mid V_{t-1}) = E(D_t \mid V_{t-1}) - (T_t^*)^{-1} E(T_t^{-1} D_t \mid V_{t-1}). \tag{G.8}$$
Plugging (G.8) into (G.7) yields

\[ \tilde{A}_t^\bullet = T_t^{-1}\{E(D_t | V_{t-1}) - (T_t^\bullet)^{-1}E(T_t^{-1}D_t | V_{t-1}) + D_t - E(D_t | V_{t-1})\} \]

\[ = -(T_tT_t^\bullet)^{-1}E(T_t^{-1}D_t | V_{t-1}) + T_t^{-1}D_t. \]

This completes the proof. \(\square\)

**Lemma G.7** (Projection onto \(\tilde{\Gamma}_t^4\)). Consider \(\tilde{\Gamma}_t^4 = \{A_t^\bullet - E(Q_tA_t^\bullet | V_{t-1})(T_t^\bullet)^{-1}T_t^{-1}W_t\sigma_{t+1} : A_t^\bullet \in \Lambda_t^\bullet\}\) as defined in Lemma F.4. For any \(h(\sigma_{t+1}, V_t) \in \mathcal{H}\), its projection onto \(\tilde{\Gamma}_t^4\) is

\[ \Pi\{h(\sigma_{t+1}, V_t) | \tilde{\Gamma}_t^4\} = O_3^* O_4^*(h) - E\{O_3^* O_4^*(h)Q_t | V_{t-1}\}W_{t,t-1}\epsilon_t, \quad \text{(G.9)} \]

where

\[ \epsilon_t = T_t^{-1}W_t\sigma_{t+1} + Q_t, \]

\[ W_{t,t-1} = \text{Var}(\epsilon_t | V_{t-1})^{-1}, \]

\[ O_3(h) = E(h | H_t) - E(h | V_{t-1}), \]

\[ O_3^*(h) = O_3(h), \]

\[ O_4(h) = h - E(hQ_t | V_{t-1})(T_t^\bullet)^{-1}T_t^{-1}W_t\sigma_{t+1}, \]

\[ O_4^*(h) = h - E\{h(T_t^\bullet)^{-1}T_t^{-1}W_t\sigma_{t+1} | V_{t-1}\}Q_t, \]

and

\[ O_3^* O_4^*(h) = E(h | H_t) - E(h | V_{t-1}) - E\{h(T_t^\bullet)^{-1}T_t^{-1}W_t\sigma_{t+1} | V_{t-1}\}Q_t. \quad \text{(G.10)} \]

In particular, if \(h = h(V_t)\), then

\[ \Pi\{h(V_t) | \tilde{\Gamma}_t^4\} = E(h | H_t) - E(h | V_{t-1}) - E(hQ_t | V_{t-1})W_{t,t-1}\epsilon_t; \quad \text{(G.11)} \]

if \(h = h(V_{t-1})\), then \(\Pi\{h(V_{t-1}) | \tilde{\Gamma}_t^4\} = 0\). Here, \(W_t, T_t, T_t^\bullet\) are defined in Lemma F.4.

**Proof.** By definition it is straightforward that \(\tilde{\Gamma}_t^4 = O_4 \circ O_3(\mathcal{H})\) and that \(O_3^* = O_3\). To derive
$O^*_q$, for any $h, g \in \mathcal{H}$ we have

$$< O_4(h), g > = E[\{h - E(hQ_t \mid V_{t-1})(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1}\}g]$$

$$= E(hg) - E\{E(hQ_t \mid V_{t-1})(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1+g}\}$$

$$= E(hg) - E[hQ_tE\{(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1+g} \mid V_{t-1}\}]$$

$$= E[h\sigma_{t+1+g} = E\{(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1+g} \mid V_{t-1}\}g],$$

so $O^*_q(h) = h - E\{h(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1} \mid V_{t-1}\}Q_t$. Using the fact that $Q_t = Q_t(H_t)$ and $E(Q_t \mid V_{t-1}) = 0$, we have

$$O^*_qO^*_q(h) = E\{O^*_q(h) \mid H_t\} - E\{O^*_q(h) \mid V_{t-1}\}$$

$$= E\{h - E\{h(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1} \mid V_{t-1}\}Q_t \mid H_t\}$$

$$- E\{h - E\{h(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1} \mid V_{t-1}\}Q_t \mid V_{t-1}\}$$

$$= E(h \mid H_t) - E(h \mid V_{t-1}) - E\{h(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1} \mid V_{t-1}\}Q_t.$$

To derive the projection, for a given $h(\sigma_{t+1}, V_t) \in \mathcal{H}$, suppose

$$\Pi\{h(\sigma_{t+1}, V_t) \mid \tilde{H}_t\} = A^*_t - E(Q_tA^*_t \mid V_{t-1})(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1} \equiv h_p$$

(G.12)

for some $A^*_t \in \mathcal{A}^*_t$. We calculate a few terms:

$$E(h_p \mid H_t) = E\{A^*_t - E(Q_tA^*_t \mid V_{t-1})(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1} \mid H_t\}$$

$$= A^*_t - E\{E(Q_tA^*_t \mid V_{t-1})(T^*_t)^{-1}T_t^{-1}W_tE(\sigma_{t+1} \mid V_t) \mid H_t\}$$

$$= A^*_t,$$

$$E(h_p \mid V_{t-1}) = E\{E(h_p \mid H_t) \mid V_{t-1}\} = E(A^*_t \mid V_{t-1}) = 0,$$

and

$$E\{h_p(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1} \mid V_{t-1}\}$$

$$= E\{A^*_t(T^*_t)^{-1}T_t^{-1}W_t\sigma_{t+1} \mid V_{t-1}\} - E\{E(Q_tA^*_t \mid V_{t-1})(T^*_t)^{-2}T_t^{-2}W_t^2\sigma_{t+1}^2 \mid V_{t-1}\}$$

$$= E\{A^*_t(T^*_t)^{-1}T_t^{-1}W_tE(\sigma_{t+1} \mid H_t) \mid V_{t-1}\} - E\{E(Q_tA^*_t \mid V_{t-1})(T^*_t)^{-2}E\{T_t^{-2}E(W_t^2\sigma_{t+1}^2 \mid H_t) \mid V_{t-1}\}$$

$$= 0 - E\{Q_tA^*_t \mid V_{t-1}\}(T^*_t)^{-2}E(T_t^{-2}T_t \mid V_{t-1})$$

$$= - E\{Q_tA^*_t \mid V_{t-1}\}(T^*_t)^{-1},$$

(G.13)
where the second to last equality in (G.13) follows from (G.11). Plugging them into (G.10) yields

\[
O_3^*O_4^*(h_p) = E(h_p \mid H_t) - E(h_p \mid V_{t-1}) - E\{h_p(T_t^\cdot)^{-1}T_t^{-1}W_t\sigma_{t+1} \mid V_{t-1}\}Q_t
= A_t^\cdot + E(Q_tA_t^\cdot \mid V_{t-1})(T_t^\cdot)^{-1}Q_t.
\]

A functional analysis result implies that \(O_3^*O_4^*(h(\sigma_{t+1},V_t)) = O_3^*O_4^*(h_p)\), i.e.,

\[
O_3^*O_4^*(h) = A_t^\cdot + E(Q_tA_t^\cdot \mid V_{t-1})(T_t^\cdot)^{-1}Q_t. \quad \text{(G.14)}
\]

Multiply both sides by \(Q_t\) then take \(E(\cdot \mid V_{t-1})\), (G.14) becomes

\[
E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\} = E(A_t^\cdot Q_t \mid V_{t-1}) + E(A_t^\cdot Q_t \mid V_{t-1})(T_t^\cdot)^{-1}E(Q_t^2 \mid V_{t-1}). \quad \text{(G.15)}
\]

Noting that \(E(Q_t \mid V_{t-1}) = 0\), (G.15) implies that

\[
E(A_t^\cdot Q_t \mid V_{t-1}) = \frac{E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}}{1 + (T_t^\cdot)^{-1}\text{Var}(Q_t \mid V_{t-1})}.
\]

Plugging into (G.14) and we have

\[
A_t^\cdot = O_3^*O_4^*(h) - \frac{E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}Q_t}{T_t^\cdot + \text{Var}(Q_t \mid V_{t-1})}.
\]

This implies that

\[
E(Q_tA_t^\cdot \mid V_{t-1}) = E\{Q_tO_3^*O_4^*(h) \mid V_{t-1}\} - \frac{E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}E(Q_t^2 \mid V_{t-1})}{T_t^\cdot + \text{Var}(Q_t \mid V_{t-1})}
= \frac{E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}T_t^\cdot}{T_t^\cdot + \text{Var}(Q_t \mid V_{t-1})}.
\]

Therefore, by the definition of \(h_p\) in (G.12) we have

\[
h_p = O_3^*O_4^*(h) - \frac{E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}Q_t}{T_t^\cdot + \text{Var}(Q_t \mid V_{t-1})}
- \frac{E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}T_t^{-1}W_t\sigma_{t+1}}{T_t^\cdot + \text{Var}(Q_t \mid V_{t-1})}
= O_3^*O_4^*(h) - E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}\frac{T_t^{-1}W_t\sigma_{t+1} + Q_t}{\text{Var}(T_t^{-1}W_t\sigma_{t+1} + Q_t \mid V_{t-1})}
= O_3^*O_4^*(h) - E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}W_{t,t-1}\epsilon_t,
\]

61
where the second to last equality follows from Lemma G.11. This proves (G.9).

If \( h = h(V_t) \), then

\[
E\{h(V_t)(T_t^{\bullet})^{-1}T_t^{-1}W_t\sigma_{t+1} \mid V_{t-1}\}Q_t
\]

so \( O_3^*O_4^*(h) = E(h \mid H_t) - E(h \mid V_{t-1}) \) and

\[
E\{O_3^*O_4^*(h)Q_t \mid V_{t-1}\}
\]

This proves (G.11). If \( h = h(V_{t-1}) \), then \( O_3^*O_4^*(h) = 0 \) and hence \( \Pi\{h(V_{t-1}) \mid \tilde{\Gamma}^4_t\} = 0 \). This completes the proof.

**Lemma G.8** (Projection onto \( \tilde{\Lambda}_t^6 \)). For any \( h \in \mathcal{H} \), \( \Pi(h \mid \tilde{\Lambda}_t^6) = E(h\epsilon_t \mid V_{t-1})W_{t,t-1}\epsilon_t \), where \( \epsilon_t = T_t^{-1}W_t\sigma_{t+1} + Q_t \) and \( W_{t,t-1} = \text{Var}(\epsilon_t \mid V_{t-1})^{-1} \).

**Proof.** Because \( \tilde{\Lambda}_t^6 = \{a(V_{t-1})\epsilon_t : a(V_{t-1})\} \) is any function \( \in \mathbb{R}^p \), we have \( E(h\epsilon_t \mid V_{t-1})W_{t,t-1}\epsilon_t \in \tilde{\Lambda}_t^6 \). So it suffices to show that for any \( a(V_{t-1})\epsilon_t \in \tilde{\Lambda}_t^6 \), \( h - E(h\epsilon_t \mid V_{t-1})W_{t,t-1}\epsilon_t \perp a(V_{t-1})\epsilon_t \).

Because \( E(\epsilon_t^2 \mid V_{t-1}) = W_{t,t-1}^{-1} \), we have

\[
E\{E(h\epsilon_t \mid V_{t-1})W_{t,t-1}\epsilon_t^2a(V_{t-1})\} = E\{E(h\epsilon_t \mid V_{t-1})W_{t,t-1}E(\epsilon_t^2 \mid V_{t-1})a(V_{t-1})\}
\]

\[
= E\{E(h\epsilon_t \mid V_{t-1})a(V_{t-1})\}
\]

\[
= E\{h\epsilon_ta(V_{t-1})\},
\]

thus

\[
E\{h - E(h\epsilon_t \mid V_{t-1})W_{t,t-1}\epsilon_t \}a(V_{t-1})\epsilon_t = 0.
\]

This completes the proof.

**Remark** 1. \( \tilde{\Lambda}_t^6 \) may not be the image of a linear operator on \( \mathcal{H} \). So instead of directly deriving the adjoing operator for \( \tilde{\Lambda}_t^6 \), the form of the projection in Lemma G.8 is obtained by first considering the projection onto the following subspace of \( \tilde{\Lambda}_t^6 \):

\[
\tilde{\Lambda}_t^6 = O_5(\mathcal{H}) = \{E(h \mid V_{t-1})\epsilon_t : h \in \mathcal{H}\}.
\]
The adjoint operator $O_5^*$ for $O_5$ can be derived as follows. For any $h, g \in \mathcal{H}$,

$$< O_5(h), g > = E\{E(h \mid V_{t-1})\epsilon_t g\} = E[E\{E(h \mid V_{t-1})\epsilon_t g \mid V_{t-1}\}] = E\{E(h \mid V_{t-1}) E(\epsilon_t g \mid V_{t-1})\} = E\{h E(\epsilon_t g \mid V_{t-1})\},$$

so $O_5^*(g) = E(\epsilon_t g \mid V_{t-1})$. Now, suppose for a given $h \in \mathcal{H}$, $\Pi(h \mid \hat{\Lambda}_t^6) = h_p(V_{t-1})\epsilon_t$ for some $h_p(V_{t-1})$ satisfying $E(h_p) = 0$. By a functional analysis result we have

$$O_5^*\{h_p(V_{t-1})\epsilon_t\} = O_5^*(h),$$

i.e.,

$$E\{\epsilon_t^2 h_p(V_{t-1}) \mid V_{t-1}\} = E(\epsilon_t h \mid V_{t-1}).$$

Since $E(\epsilon_t^2 \mid V_{t-1}) = W_{t,t-1}^{-1}$, the above display implies $h_p(V_{t-1}) = W_{t,t-1} E(\epsilon_t h \mid V_{t-1})$, and thus $\Pi(h \mid \hat{\Lambda}_t^6) = E(\epsilon_t h \mid V_{t-1}) W_{t,t-1} \epsilon_t$. Having obtained this, we then verified by definition that it is also the projection onto $\hat{\Lambda}_t^6$ in the proof of Lemma G.8.

**Lemma G.9** (Projection of $h(\sigma_{t+1}, V_t)$ onto $\Lambda_t^+$). For any $B = b(V_{t+1}) \in \mathcal{H}$, let $h(\sigma_{t+1}, V_t) = E(B \mid \sigma_{t+1}, V_t) = E(B \mid \sigma_{t+1}, V_t) - E(B \mid V_t)$. Then the projection of $h(\sigma_{t+1}, V_t)$ onto $\Lambda_t^+$ is

$$\Pi\{h(\sigma_{t+1}, V_t) \mid \Lambda_t^+\} = \{R_t - T_t^{-1} E(R_t W_t \mid X_t)\} W_t \sigma_{t+1},$$

where $R_t = E(B \sigma_{t+1} \mid V_t)$, and $W_t, T_t$ are defined in Lemma F.4.

**Proof.** By Lemma F.4 and Lemma G.12 to compute $\Pi\{h(\sigma_{t+1}, V_t) \mid \Lambda_t^+\}$ it suffices to calculate sequentially the projection of $h(\sigma_{t+1}, V_t)$ onto $\Lambda_t^{1,1}, \Lambda_t^{2,1}, (\bigoplus_{m=1}^{t-1} \Gamma_m^3)^1, (\bigoplus_{m=1}^{t-1} \Lambda_m^*)^1, \tilde{\Lambda}_t^{5,1}, \tilde{\Gamma}_t^{4,1}$ and $\tilde{\Lambda}_t^{6,1}$.

(i) For any $A_t^1(V_{t+1}) \in \Lambda_t^1$, $E\{h(\sigma_{t+1}, V_t) A_t^1\} = E\{h(\sigma_{t+1}, V_t) E(A_t^1 \mid V_t, Y_{t+1})\} = 0$. So $h(\sigma_{t+1}, V_t) \in \Lambda_t^{1,1}$ and $\Pi(h \mid \Lambda_t^{1,1}) = h$.  

63
(ii) By Lemma A.4 and the fact that $E(h | V_t) = 0$, we have $\Pi(h | \Lambda^2_t) = h - E(h \sigma_{t+1} | V_t)W_t\sigma_{t+1}$. Note that

$$E(h \sigma_{t+1} | V_t) = E\{E(B | \sigma_{t+1}, V_t)\sigma_{t+1} | V_t\} - E\{E(B | V_t)\sigma_{t+1} | V_t\}$$

$$= E(B \sigma_{t+1} | V_t) - 0 = R_t,$$

so we have

$$\Pi(h | \Lambda^2_t) = h - \Pi(h | \Lambda^2_t) = R_t W_t \sigma_{t+1}.$$

(iii) For any $g(V_t) \in \mathcal{H}$, we have $E\{g(V_t) R_t W_t \sigma_{t+1}\} = E\{g(V_t) R_t W_t E(\sigma_{t+1} | V_t)\} = 0$, so $R_t W_t \sigma_{t+1} \in (\bigoplus_{m=1}^{t} \Gamma^3_m)^\perp$ and $R_t W_t \sigma_{t+1} \in (\bigoplus_{m=1}^{t-1} \Lambda^*_{m})^\perp$. Therefore, $\Pi\{R_t W_t \sigma_{t+1} | (\bigoplus_{m=1}^{t} \Gamma^3_m)^\perp\} = R_t W_t \sigma_{t+1}$ and $\Pi\{R_t W_t \sigma_{t+1} | (\bigoplus_{m=1}^{t-1} \Lambda^*_{m})^\perp\} = R_t W_t \sigma_{t+1}$.

(vi) To use Lemma A.6 to compute $\Pi(R_t W_t \sigma_{t+1} | \tilde{\Lambda}^5_t)$, we first calculate a few terms:

$$D_t = E(R_t W_t^2 \sigma_{t+1}^2 | H_t) = E\{R_t W_t^2 E(\sigma_{t+1}^2 | V_t) | H_t\} = E(R_t W_t | H_t),$$

$$\tilde{\Lambda}^*_t = -E(D_t T^{-1}_t | V_{t-1})(T^*_t T^{-1}_t) + D_t T^{-1}_t$$

$$= -E\{E(R_t W_t | H_t) T^{-1}_t | V_{t-1}\}(T^*_t T^{-1}_t) + E(R_t W_t | H_t) T^{-1}_t$$

$$= -E(R_t W_t T^{-1}_t | V_{t-1})(T^*_t T^{-1}_t) + E(R_t W_t | H_t) T^{-1}_t$$

$$= -R_{t-1} (T^*_t) T^{-1}_t + E(R_t W_t | H_t) T^{-1}_t,$$

where we define $R_{t-1} = E(R_t W_t T^{-1}_t | V_{t-1})$. So by Lemma A.6 we have

$$\Pi(R_t W_t \sigma_{t+1} | \tilde{\Lambda}^5_t) = \tilde{\Lambda}^*_t W_t \sigma_{t+1} = \{E(R_t W_t | H_t) - R_{t-1} (T^*_t) T^{-1}_t\} T^{-1}_t W_t \sigma_{t+1},$$

and

$$\Pi(R_t W_t \sigma_{t+1} | \tilde{\Lambda}^5_{t-1}) = R_t W_t \sigma_{t+1} - \Pi(R_t W_t \sigma_{t+1} | \tilde{\Lambda}^5_t)$$

$$= \{R_t T_t + R_{t-1} (T^*_t) T^{-1}_t - E(R_t W_t | H_t)\} T^{-1}_t W_t \sigma_{t+1}$$

$$\equiv h_1(V_t) T^{-1}_t W_t \sigma_{t+1},$$

where we define $h_1(V_t) = R_t T_t + R_{t-1} (T^*_t) T^{-1}_t - E(R_t W_t | H_t)$. 

64
(v) Now we will use Lemma G.7 to compute \( \Pi \{ h_1(V_t)T_t^{-1}W_t\sigma_{t+1} | \tilde{\Gamma}_{t}^{4,4} \} \). Since \( E(\sigma_{t+1} | V_t) = 0 \), we have

\[
E\{ h_1(V_t)T_t^{-1}W_t\sigma_{t+1} | H_t \} = E\{ h_1(V_t)T_t^{-1}W_t\sigma_{t+1} | V_{t-1} \} = 0.
\]

We also have (using \( E(\sigma_{t+1}^2 | V_t) = W_t^{-1} \))

\[
E\{ h_1(V_t)T_t^{-1}W_t\sigma_{t+1} \times (T_t^\bullet)^{-1}T_t^{-1}W_t\sigma_{t+1} | V_{t-1} \}
= E\{ h_1(V_t)(T_t^\bullet)^{-1}T_t^{-2}W_t\sigma_{t+1} | V_{t-1} \}
= E\{ h_1(V_t)(T_t^\bullet)^{-1}T_t^{-2}W_t | V_{t-1} \}
= E\{ R_t(T_t^\bullet)^{-1}T_t^{-2}W_t | V_{t-1} \} + E\{ R_{t-1}(T_t^\bullet)^{-2}T_t^{-2}W_t | V_{t-1} \}
- E\{ (R_tW_t | H_t)(T_t^\bullet)^{-1}T_t^{-2}W_t | V_{t-1} \}.
\] (G.16)

We compute out each term in (G.16):

\[
E\{ R_t(T_t^\bullet)^{-1}T_t^{-1}W_t | V_{t-1} \} = E\{ R_tT_t^{-1}W_t | V_{t-1} \} (T_t^\bullet)^{-1} = R_{t-1}(T_t^\bullet)^{-1},
\]

\[
E\{ R_{t-1}(T_t^\bullet)^{-2}T_t^{-2}W_t | V_{t-1} \} = R_{t-1}(T_t^\bullet)^{-2}E\{ T_t^{-2}W_t | V_{t-1} \}
= R_{t-1}(T_t^\bullet)^{-2}E\{ T_t^{-2}E(W_t | H_t) | V_{t-1} \}
= R_{t-1}(T_t^\bullet)^{-2}E\{ T_t^{-1} | V_{t-1} \} = R_{t-1}(T_t^\bullet)^{-1},
\]

\[
E\{ (R_tW_t | H_t)(T_t^\bullet)^{-1}T_t^{-2}W_t | V_{t-1} \} = E\{ (R_tW_tT_t^{-2} | H_t)E(W_t | H_t) | V_{t-1} \} (T_t^\bullet)^{-1}
= E\{ (R_tW_tT_t^{-1} | H_t) | V_{t-1} \} (T_t^\bullet)^{-1}
= R_{t-1}(T_t^\bullet)^{-1}.
\]

Hence, (G.16) becomes

\[
E\{ h_1(V_t)T_t^{-1}W_t\sigma_{t+1} \times (T_t^\bullet)^{-1}T_t^{-1}W_t\sigma_{t+1} | V_{t-1} \} = R_{t-1}(T_t^\bullet)^{-1}.
\]

By the definition of \( O_3^\ast \) and \( O_4^\ast \) in Lemma G.7 we have

\[
O_3^\ast O_4^\ast \{ h_1(V_t)T_t^{-1}W_t\sigma_{t+1} \} = -R_{t-1}(T_t^\bullet)^{-1}Q_t,
\]

and

\[
E[O_3^\ast O_4^\ast \{ h_1(V_t)T_t^{-1}W_t\sigma_{t+1} \}Q_t | V_{t-1}] = -R_{t-1}(T_t^\bullet)^{-1} \text{Var}(Q_t | V_{t-1}).
\]
With the above computation, Lemma \textbf{G.7} implies that

\[
\Pi\{h_1(V_t)T_t^{-1}W_t\sigma_{t+1} | \tilde{\Gamma}_t^4\} \\
= O_1^tO_1^t\{h_1(V_t)T_t^{-1}W_t\sigma_{t+1}\} - E[O_1^tO_1^t\{h_1(V_t)T_t^{-1}W_t\sigma_{t+1}\}Q_t | V_{t-1}]W_{t,t-1}\epsilon_t \\
= -R_{t-1}(T_t^*)^{-1}Q_t + R_{t-1}(T_t^*)^{-1}\text{Var}(Q_t | V_{t-1})W_{t,t-1}\epsilon_t.
\]

Thus, the projection \(\Pi\{h_1(V_t)T_t^{-1}W_t\sigma_{t+1} | \tilde{\Gamma}_t^{4,\perp}\}\) equals

\[
\Pi\{h_1(V_t)T_t^{-1}W_t\sigma_{t+1} | \tilde{\Gamma}_t^{4,\perp}\} = h_1(V_t)T_t^{-1}W_t\sigma_{t+1} - \Pi\{h_1(V_t)T_t^{-1}W_t\sigma_{t+1} | \tilde{\Gamma}_t^4\} \\
= \{R_tT_t + R_{t-1}(T_t^*)^{-1} - E(R_tW_t | H_t)\}T_t^{-1}W_t\sigma_{t+1} \\
+ R_{t-1}(T_t^*)^{-1}Q_t - R_{t-1}(T_t^*)^{-1}\text{Var}(Q_t | V_{t-1})W_{t,t-1}\epsilon_t \\
= \{R_t - T_t^{-1}E(R_tW_t | H_t)\}W_t\sigma_{t+1} + R_{t-1}(T_t^*)^{-1}\epsilon_t - R_{t-1}(T_t^*)^{-1}\text{Var}(Q_t | V_{t-1})W_{t,t-1}\epsilon_t \\
= \{R_t - T_t^{-1}E(R_tW_t | H_t)\}W_t\sigma_{t+1} + R_{t-1}W_{t,t-1}\epsilon_t,
\]

where the last equality follows from Lemma \textbf{G.11}.

(vi) Denote by \(h_2(\sigma_{t+1}, V_t) = \{R_t - T_t^{-1}E(R_tW_t | H_t)\}W_t\sigma_{t+1}\) and \(h_3(\sigma_{t+1}, V_t) = R_{t-1}W_{t,t-1}\epsilon_t\). Now we will use Lemma \textbf{G.8} to compute \(\Pi(h_2 + h_3 | \tilde{\Lambda}_t^{6,\perp})\). We first calculate a few terms:

\[
E\{h_2(\sigma_{t+1}, V_t)\epsilon_t | V_{t-1}\} = E(R_tW_t\sigma_{t+1}\epsilon_t | V_{t-1}) - E\{E(R_tW_t | H_t)T_t^{-1}W_t\sigma_{t+1}\epsilon_t | V_{t-1}\}. \\
\text{(G.17)}
\]

Using the fact that \(E(\sigma_{t+1} | V_t) = 0\) and \(E(\sigma_{t+1}^2 | V_t) = W_t^{-1}\), we have. By Lemma \textbf{G.11}(iv) we have

\[
E(R_tW_t\sigma_{t+1}\epsilon_t | V_{t-1}) = E\{R_tW_tE(\sigma_{t+1}\epsilon_t | V_t) | V_{t-1}\} = E(R_tW_tT_t^{-1} | V_{t-1}) = R_{t-1},
\]

\[
E\{E(R_tW_t | H_t)T_t^{-1}W_t\sigma_{t+1}\epsilon_t | V_{t-1}\} = E(R_tW_tT_t^{-1} | V_{t-1}) = R_{t-1},
\]

66
Lemma G.10 (Projection of $h(H_t)$ onto $\Lambda_t^\perp$). For any $B = b(V_{T+1}) \in \mathcal{H}$, let $h(H_t) = E(B \mid H_t) - E(B \mid V_{t-1})$. Then $h(H_t) \in \Lambda_t$, i.e., $\Pi\{h(H_t) \mid \Lambda_t^\perp\} = 0$.

Proof. By Lemma F.4 and Lemma G.12, to compute $\Pi\{h(H_t) \mid \Lambda_t^\perp\}$ it suffices to calculate sequentially the projection of $h(H_t)$ onto $\Lambda_t^{1,\perp}, \Lambda_t^{2,\perp}, (\bigoplus_{m=1}^t \Gamma_m)^\perp, (\bigoplus_{m=1}^{t-1} \Lambda_m)^\perp, \tilde{\Lambda}_t^{5,\perp}, \tilde{\Gamma}_t^{4,\perp}$ and $\tilde{\Lambda}_t^{6,\perp}$.

(i) For any $A_t^1(V_{T+1}) \in \Lambda_t^1$, $E\{h(H_t)A_t^1\} = E\{h(H_t)E(A_t^1 \mid V_t, Y_{t+1})\} = 0$. So $h(H_t) \in \Lambda_t^{1,\perp}$ and $\Pi(h \mid \Lambda_t^{1,\perp}) = h$.

(ii) For any $A_t^2(\sigma_{t+1}, V_t) \in \Lambda_t^2$, $E\{h(H_t)A_t^2\} = E\{h(H_t)E(A_t^2 \mid V_t)\} = 0$. So $h(H_t) \in \Lambda_t^{2,\perp}$ and $\Pi(h \mid \Lambda_t^{2,\perp}) = h$. 

This completes the proof.
(iii) For any \( A_m^3(V_m) \in \Gamma_m^3 \) with \( 1 \leq m \leq t - 1 \), we have \( E\{h(H_t)A_m^3(V_m)\} = E\{h(H_t) | V_{t-1}\}A_m^3(V_m) = 0 \). For any \( A_t^3(V_t) \in \Gamma_t^3 \), we have \( E\{h(H_t)A_t^3(V_t)\} = E[h(H_t)E\{A_m^3(V_m) | H_t\}] = 0 \). So \( h(H_t) \in (\bigoplus_{m=1}^t \Gamma_m^3)^\perp \) and \( \Pi\{h | (\bigoplus_{m=1}^t \Gamma_m^3)^\perp\} = h \).

(iv) For any \( A_m^•(H_m) \in \Lambda_m^• \) with \( 1 \leq m \leq t - 1 \), we have \( E\{h(H_t)A_m^•(H_m)\} = E[h(H_t) | V_{t-1}\}A_m^•(H_m) = 0 \). So \( h(H_t) \in (\bigoplus_{m=1}^{t-1} \Lambda_m^•)^\perp \) and \( \Pi\{h | (\bigoplus_{m=1}^{t-1} \Lambda_m^•)^\perp\} = h \).

(v) We have \( D_t = E\{h(H_t)W_t\sigma_{t+1} | H_t\} = E\{h(H_t)W_tE(\sigma_{t+1} | V_t) | H_t\} = 0 \), so by Lemma G.6, \( \Pi\{h(H_t) | \tilde{\Lambda}^5_t\} = 0 \) and \( \Pi\{h(H_t) | \tilde{\Lambda}^{5,\perp}_t\} = h(H_t) \).

(vi) By Lemma G.7 using the fact that \( E(Q_t | V_{t-1}) = 0 \), we have

\[
\Pi\{h(H_t) | \tilde{\Gamma}^4_t\} = E(h | H_t) - E(h | V_{t-1}) - E(hQ_t | V_{t-1})W_{t,t-1}\epsilon_t \\
= h(H_t) - E\{E(B | H_t)Q_t - E(B | V_{t-1})Q_t | V_{t-1}\}W_{t,t-1}\epsilon_t \\
= h(H_t) - E(BQ_t | V_{t-1})W_{t,t-1}\epsilon_t,
\]

so \( \Pi\{h(H_t) | \tilde{\Gamma}^{4,\perp}_t\} = h(H_t) - \Pi\{h(H_t) | \tilde{\Gamma}^4_t\} = E(BQ_t | V_{t-1})W_{t,t-1}\epsilon_t \).

(vii) By definition we have \( E(BQ_t | V_{t-1})W_{t,t-1}\epsilon_t \in \tilde{\Lambda}^6_t \), so \( \Pi\{E(BQ_t | V_{t-1})W_{t,t-1}\epsilon_t | \tilde{\Lambda}^{6,\perp}_t\} = 0 \). This completes the proof.

Lemma G.11. With \( W_t = \text{Var}(\sigma_{t+1} | V_t)^{-1} \), \( T_t = E(W_t | H_t) \), \( T_t^• = E(T_t^{-1} | V_{t-1}) \), \( W_{t,t-1} = \text{Var}(T_t^{-1}W_t\sigma_{t+1} + Q_t | V_{t-1})^{-1} \), we have

(i) \( E(W_t^2\sigma_{t+1}^2 | H_t) = T_t \).

(ii) \( W_{t,t-1}^{-1} = \text{Var}(T_t^{-1}W_t\sigma_{t+1} | V_{t-1}) + \text{Var}(Q_t | V_{t-1}) = T_t^• + \text{Var}(Q_t | V_{t-1}) \).

(iii) \( 1 - \text{Var}(Q_t | V_{t-1})W_{t,t-1} = T_t^• \).

(iv) \( E(\sigma_{t+1}\epsilon_t | V_t) = T_t^{-1} \).
Proof. For (i), because $E(\sigma_t + 1 \mid V_t) = 0$, we have

$$E(W_t^2 \sigma_t^2 \mid H_t) = E\{E(\sigma_t + 1 \mid V_t)^2 \mid H_t\}$$
$$= E[E\{E(\sigma_t + 1 \mid V_t)^2 \mid V_t\} \mid H_t]$$
$$= E\{E(\sigma_t + 1 \mid V_t)^2 \mid H_t\} = T_t.$$

For (ii), we have

$$\text{Var}(T_t^{-1} W_t \sigma_t + Q_t \mid V_t - 1) = E\{(T_t^{-1} W_t \sigma_t + Q_t)^2 \mid V_t - 1\}$$
$$= E(T_t^{-2} W_t^2 \sigma_t^2 \mid V_t - 1) + E(Q_t^2 \mid V_t - 1)$$
$$= E(T_t^{-1} \mid V_t - 1) + \text{Var}(Q_t \mid V_t - 1) = T_t^* + \text{Var}(Q_t \mid V_t - 1).$$

(iii) is an immediate implication of (ii).

For (iv), we have

$$E(\sigma_t + 1 \epsilon_t \mid V_t) = E\{\sigma_t + 1 (T_t^{-1} W_t \sigma_t + Q_t) \mid V_t\}$$
$$= T_t^{-1} W_t E(\sigma_t^2 \mid V_t) + E(\sigma_t Q_t \mid V_t)$$
$$= T_t^{-1} W_t W_t^{-1} + 0$$
$$= T_t^{-1}.$$ 

This completes the proof.

Lemma G.12. Suppose $\Lambda_1$ and $\Lambda_2$ are two subspaces of the Hilbert space $\mathcal{H}$, and they are orthogonal to each other. Then for any $h \in \mathcal{H}$, we have

$$\Pi\{h \mid (\Lambda_1 \oplus \Lambda_2)^\perp\} = \Pi\{\Pi(h \mid \Lambda_1^\perp) \mid \Lambda_2^\perp\}.$$ 

Proof. This is a standard Hilbert space result. See, for example, Akhiezer & Glazman (2013).