Off-policy Policy Evaluation Under Unobserved Confounding

Anonymous Authors[1]

Abstract

When observed decisions depend only on observed features, off-policy policy evaluation (OPE) methods for sequential decision making problems allow evaluating the performance of evaluation policies before deploying them. This assumption is often violated due to the presence of unobserved confounders, variables that impact both the decisions and their outcomes. We assess the robustness of OPE methods by developing worst-case bounds on the performance of a evaluation policy under different models of confounding. When unobserved confounders can affect every decision in an episode, we demonstrate that even small amounts of per-decision confounding can heavily bias OPE methods. Fortunately, in a number of important settings found in healthcare, policy-making, operations, and technology, unobserved confounders may primarily affect only one of the many decisions made. Under this less pessimistic model of one-decision confounding, we propose an efficient loss-minimization-based procedure for computing worst-case bounds on OPE estimates, and prove its statistical consistency. On simulated healthcare examples, we demonstrate that our method allows reliable off-policy evaluation by invalidating non-robust results, and providing certificates of robustness.

1. Introduction

New technology and regulatory shifts allow collection of unprecedented amounts of data on past decisions and their associated outcomes, ranging from product recommendation systems to medical treatment decisions. This presents unique opportunities to leverage off-policy observational data to inform better decision-making. When online experimentation is expensive or risky, it is crucial to leverage prior data to evaluate the performance of a sequential decision policy before deploying it. While epidemiology has long been interested in dynamic treatment regime estimation, the reinforcement learning community is increasingly paying attention to batch reinforcement learning (RL) because of new models and data availability (see e.g. (Thomas et al., 2019; Liu et al., 2018; Le et al., 2019; Thomas et al., 2015; Komorowski et al., 2018b; Hanna et al., 2017; Gottesman et al., 2019a;b)). We focus on the common scenario where decisions are made in episodes by an unknown behavioral policy, each involving a sequence of decisions.

A central challenge in OPE is that the estimand is inherently a counterfactual quantity: what would the resulting outcomes be *if an alternate policy had been used* (the counterfactual) instead of behavior policy used in the collected data (the factual). As a result, OPE requires causal reasoning about whether the decisions caused observed differences in rewards, as opposed to being caused by some unobserved confounding variable that simultaneously affect both observed decisions and the states or rewards (Hernán and Robins, 2020; Pearl, 2009).

In order to make counterfactual evaluations possible, a standard assumption—albeit often overlooked and unstated—is to require that the behavior policy also does not depend on any unobserved/latent variables that also affect the future states or rewards (no unobserved confounding). We refer to this assumption as *sequential ignorability*, following the line of works on dynamic treatment regimes (Robins, 1986; 1997; Murphy et al., 2001; Murphy, 2003). Sequential ignorability, however, is frequently violated in observational problems where the behavioral policy is unknown. In healthcare, business operations, and even automated systems in tech, decisions are often made with respect to unlogged data correlated with future potential outcomes.

In this work, we develop and analyze a framework that can quantify the impact of unobserved confounders on off-policy policy evaluations, providing certificates of robustness. Since OPE is generally impossible under arbitrary amounts of unobserved confounding, we begin by positing a model that explicitly limits their influence on decisions. In Section 4, we illustrate that when unobserved confounders can affect all decisions, even small amounts of confounding can have an exponential (in the number of decision steps)

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

impact on the error of the resulting off policy evaluation. In this sense, the validity of OPE can almost always be questioned under presence of unobserved confounding that affect all time steps. Fortunately, in a number of important applications, unobserved confounders may only affect a single decision, particularly in scenarios where experts are a high level decision-maker that use unrecorded information to make a initial decision, after which a standard set of protocols are followed based on well-recorded observations.

Under our less pessimistic model of single-decision confounding, we develop bounds on the expected rewards of the evaluation policy (Section 5). We use functional convex duality to derive a dual relaxation, and show that it can be computed by solving a loss minimization problem. Our procedure allows analyzing sensitivity of OPE methods to confounding in realistic scenarios involving continuous state and rewards, over a potentially large horizon. We prove that an empirical approximation of our procedure is consistent, allowing estimation from observed past decisions.

On simulation examples of dynamic treatment regimes for autism and sepsis management, we illustrate how our single-decision confounding model allows us to obtain informative bounds over meaningful amounts of potential confounding. Our approach can both provide a certificate of the robustness of OPE under a certain amount of unobserved confounding, as well as identify when bias in OPE can raise concerns for validity of selecting the best policy among a set of candidates. As we illustrate, developing tools for a meaningful sensitivity analysis is nontrivial: a naive bound yields prohibitively conservative estimates that almost lose robustness certificates for even neglible amounts of confounding, whereas our loss-minimization-based bounds on policy values is informative.

1.1. Motivating example: managing sepsis patients

Managing sepsis in ICU patients is an extremely important problem, accountable for 1/3 of deaths in hospitals (Howell and Davis, 2017). Sepsis treatment decisions are made by a clinical care team, including nurses, residents, and ICU attending physicians and specialists (Rhodes et al., 2017). Difficulties of care often lead to making decisions based off of imperfect information, leading to substantial room for improvement. AI-based approaches provide an opportunity for optimal automated management of medications, freeing the care team to allocate more resources to critical cases. Automated approaches can manage important medications for sepsis, including antibiotics and vasopressors, and decide to notify the care team about when a patient should be placed on a mechanical ventilator. Motivated by these opportunities, and the availability of ICU data from MIMIC-3 (Johnson et al., 2016), several AI-based approaches for sepsis management system have been proposed (Futoma et al., 2018; Komorowski et al., 2018a; Raghu et al., 2017). Due to safety concerns, new treatment policies need to be evaluated offline before a more thorough validation. Confounding, however, is a serious issue in data generated from an ICU. Patients in emergency departments often do not have an existing record in the hospital's electronic health system, leaving a substantial amount of patient-specific information unobserved in subsequent offline analysis. As a prominent example, comorbidities that significantly complicate the cases of sepsis (Brent, 2017) are often unrecorded. Private communication with an emergency department physician revealed that *initial* treatment of antibiotics at admission to the hospital are often confounded by unrecorded factors that affect the eventual outcome (death or discharge from the ICU). For example, comorbidities such as undiagnosed or improperly heart failure can delay diagnosis of sepsis, leading to slower implementation of antibiotic treatments. More generally, there is considerable discussion in the medical literature on the importance of quickly beginning antibiotic treatment, with frequently noted concerns about confounding, as these discussions are largely based on off-policy observational data collected from ICUs (Seymour et al., 2017; Sterling et al., 2015). Antibiotics are of particular interest given the recent debate regarding the importance of early treatment, and the risks of over-prescription.

We consider a scenario where one wishes to evaluate between two automated policies: optimal treatment policies that differ only in initially *avoiding*, or *prescribing* antibiotics. The latter is often considered a better treatment for sepsis, as it is caused by an infection. In this example, unobserved factors most critically effect the first decision on prescribing antibiotics upon arrival; since the care team is highly trained for treating sepsis, we assume they follow standard protocols based on observed vitals signs and lab measurements in subsequent time steps. In what follows, we assess the impact of confounding factors discussed above on OPE of automated policies, and provide certificates of robustness that guarantee gains over existing policies.

2. Related Work

Most methods for OPE for batch reinforcement learning largely rely (implicitly or explicitly) on sequential ignorability. There is an extensive body of work for off policy evaluation and optimization under this assumption, including doubly robust methods (Jiang and Li, 2015; Thomas and Brunskill, 2016) and recent work that provides semiparametric efficiency bounds (Kallus and Uehara, 2019). Often the probabilities of observing particular decisions (the behavior policy) are assumed to be known, though prior work has highlighted how errors in these quantities can bias value estimates (Liu et al., 2018) or provided estimators that learn and leverage a predictor of the decision probabilities (Nie et al., 2019; Hanna et al., 2019). Unfortunately doubly robust estimators suffer from the same bias when sequential ignorability doesn't hold, since neither the outcome model nor the importance sampling weights can correct for the effect of the unobserved confounder. The do-calculus and its sequential backdoor criterion on the associated directed acyclic graph (Pearl, 2009) also gives identification results for OPE. Like sequential ignorability, this preclude the existence of unobserved confounding variables. Therefore, methods assuming the sequential backdoor criterion holds will be biased in their presence.

The focus of this work is to study how unobserved confounding affects OPE in sequential decision making problems, and derive bounds on the evaluation policy performance in the presence of confounding. Zhang and Bareinboim (2019) derived partial identification bounds on policy performance without making model assumptions about the unobserved confounder, similar to work by Manski (1990) on bounding treatment effects. Robins et al. (2000); Robins (2004); Brumback et al. (2004) instead posit a model for how the confounding bias in each time step affects the outcome of interest and derive bounds under this model motivated by potential confounding in the analysis the effects of dynamic treatment regimes for HIV therapy on CD4 counts in HIV-positive men. Our work is complementary to these in that we instead assume a model for how the unobserved confounder affects the behavior policy, motivated by the nature of confounding in the management of sepsis patients and developmental interventions for autistic children.

For single decision making problems, a variety of methods developed in the econometrics, statistics, and epidemiology literature estimate bounds on treatment effects and mean potential outcomes based on a model for the effect of the unobserved confounder on the behavior policy (Cornfield et al., 1959; Rosenbaum and Rubin, 1983; Robins et al., 2000; Imbens, 2003; Brumback et al., 2004). Recent work has extended this to heterogeneous treatment effect estimates (Yadlowsky et al., 2018; Kallus et al., 2018) closely related to policy evaluation, and policy optimization (Kallus and Zhou, 2018). Our model is closely related to these, and naturally extends these approaches to sequential decision making (see Section 4 for a detailed discussion).

3. Formulation

Notation conventions vary substantially in the diverse set of communities interested in learning from observational data gathered on sequences of decisions and their outcomes. In this paper, we use the potential outcomes notation to make explicit which action we wish to evaluate versus which action was actually observed. In this setting, we imagine that all counterfactual (potential) states and rewards exist, but we only observe the one corresponding to the action taken (also known as partial feedback). In batch off policy reinforcement learning sequential ignorability (as we discuss further) is almost always assumed, in which case the distribution of states and rewards conditional on taking an action are equivalent to the potential outcome evaluated at that action. However, since our aim is to consider the impact of potential confounding, clarifying the difference between factual and counterfactual states and rewards is important.

We focus on domains modeled by episodic stochastic decision processes with a discrete set of actions. Let $\mathcal{A}_t$ be a finite action set of actions available at time $t = 1, .., T$. Denote a sequence of actions $a_1 \in \mathcal{A}_1, .., a_T \in \mathcal{A}_T$ by $a_{1:T}$ (and similarly $a_{t:t'}$ for arbitrary indices $1 \leq t \leq t' \leq T$, with the convention $a_{1:0} = \varnothing$). For any sequence of actions $a_{1:T}$, let $S_t(a_{1:t-1})$ and $R_t(a_{1:t})$ be the state and reward at time t: note in general there are many potential realizable states for a particular prior sequence of actions. $Y(a_{1:T}) := \sum_{t=1}^{T} \gamma^{t-1} R_t(a_{1:t})$ is the corresponding discounted sum of rewards. We denote by $W(a_{1:T}) = (S_1(a_1), .., S_T(a_{1:T-1}), R_1(a_1), .., R_T(a_{1:T}))$ all potential outcomes (over rewards and states) associated with the action sequence $a_{1:T}$. Any sum $\sum_{a_{1:t}}$ over action sequences is taken over all $a_{1:T} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_T$.

In the off-policy setting, we observe sequences of actions $A_1, .., A_T$ generated by an unknown behavioral policy $\pi_1, .., \pi_T$. Let H_t denote the observed history until time t, so that $H_1 := S_1$, and for $t = 2, .., T$, $H_t := (S_1, A_1, S_2(A_1), A_2, .., S_t(A_{1:t-1}))$. As a notational shorthand, for any fixed sequence of actions $a_{1:T}$, we denote an instantiation of the observed history following the action sequence by $H_t(a_{1:t-1})$, so that $H_1(a_{1:0}) := H_1 = S_1$, and for $t = 2, .., T$, $H_t(a_{1:t-1}) = (S_1, A_1 = a_1, S_2(a_1), .., A_{t-1} = a_{t-1}, S_t(a_{1:t-1}))$. We denote by $\mathcal{H}_t$ the set that this history takes values over.

When there is no unobserved confounding, $A_t \sim \pi_t(\cdot \mid H_t)$ since actions are generated conditional on the history H_t. When there is unobserved confounding U_t, the behavioral policy draws actions $A_t \sim \pi_t(\cdot \mid H_t, U_t)$, and we denote by $\pi_t(\cdot \mid H_t)$ the conditional distribution of A_t given only the observed history H_t, meaning we marginalize out the U_t dependence. For simplicity, we assume that previously observed rewards are included in the states, so $R_s(A_{1:s})$ is deterministic given H_t, the history and previous action We define $Y_t(a_t) := Y(A_{1:t-1}, a_t, A_{t+1:T})$ as a shorthand: semantically this means the sum of rewards which matches a trajectory of executed actions on all but one action, where on time step t action a_t is taken. Note that since a_t may not be identical to the taken action A_t, and the resulting expression for Y represents a potential outcome.

Our goal is to reliably bound the bias of evaluating the performance of a evaluation policy $\bar{\pi}_1, .., \bar{\pi}_T$ in a confounded multi-decision off-policy environment. In standard batch RL notation, this would mean we wish to bound the bias of estimating $V^{\bar{\pi}}$ using behavioral data generated under the presence of confounding variables. Let $\bar{A}_t \sim \bar{\pi}_t(\cdot \mid \bar{H}_t)$ be the actions generated by the evaluation policy at time t, where we use $\bar{H}_t := (S_1, \bar{A}_1, S_2(\bar{A}_1), \bar{A}_2, .., S_t(\bar{A}_{1:t-1}))$ and $\bar{H}_t(a_{1:t-1}) := (S_1, \bar{A}_1 = a_1, S_2(a_1), \bar{A}_2 = a_2, .., S_t(a_{1:t-1}))$ to denote the history under the evaluation policy, analogously to the shorthands $H_t, H_t(a_{1:t-1})$. We are interested in statistical estimation of the expected cumulative reward $\mathbb{E}[Y(\bar{A}_{1:T})]$ under the evaluation policy, which we call the *performance* of the evaluation policy (aka $V^{\bar{\pi}}$ in batch RL). Throughout, we assume that for all t and a_t, and almost every H_t, $\pi_t(a_t \mid H_t) > 0$ whenever $\bar{\pi}_t(a_t \mid \bar{H}_t) > 0$.

We can now state the sequential ignorability in terms of the relationship between actions and potential outcomes.

Definition 1 (Sequential Ignorability)**.** *A policy π satisfies sequential ignorability (see e.g (Robins, 1986; 2004; Murphy, 2003)) if for all $t = 1, .., T$, conditional on the history H_t generated by the policy π, $A_t \sim \pi_t(\cdot \mid H_t)$ is independent of the potential outcomes $R_t(a_{1:t}), S_{t+1}(a_{1:t}), R_{t+1}(a_{1:t+1}), S_{t+2}(a_{1:t+1}), ..,$ $S_T(a_{1:t-1}), R_T(a_{1:T})$ for all $a_{1:T} \in \mathcal{A}_1 \times \cdots \mathcal{A}_T$.*

Sequential ignorability is a natural condition required for the evaluation policy to be well-defined: any additional randomization used by the evaluation policy $\bar{\pi}_t(\cdot \mid \bar{H}_t)$ cannot depend on unobserved confounders. We assume that the evaluation policy always satisfies this assumption.

Assumption A. *The evaluation policy satisfies sequential ignorability (Definition 1).*

Off-policy policy evaluation fundamentally requires counterfactual reasoning since we only observe the state evolution $S_t(A_{1:t-1})$ and rewards $R_t(A_{1:t})$ corresponding to the actions made by the behavioral policy. The canonical assumption in batch off-policy reinforcement learning is that sequential ignorability *holds for the behavior policy*. We now briefly review how this allows identification (and thus, accurate estimation) of $\mathbb{E}[Y(\bar{A}_{1:T})]$.

Because we only observe potential outcomes $W(A_{1:t})$ evaluated at the actions $A_{1:t}$ taken by the behavior policy π_t, we need to express $\mathbb{E}[Y(\bar{A}_{1:T})]$ in terms of observable data generated by the behavioral policy π_t. Sequential ignorability of both the behavior policy and evaluation policy allows such counterfactual reasoning. The following identity is standard (see the appendix for its derivation).

Lemma 1. *Assume sequential ignorability (Definition 1) holds for both behavioral and evaluation policy. Then,*

$$\mathbb{E}[Y(\bar{A}_{1:T})] = \mathbb{E}\left[Y(A_{1:T}) \prod_{t=1}^{T} \frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}))}{\pi_t(A_t \mid H_t)}\right]$$

The RHS is called the importance sampling formula. To ease notation, we write

$$\rho_t := \frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}))}{\pi_t(A_t \mid H_t)}. \tag{1}$$

4. Bounds under unobserved confounding

Despite the advantageous implications, it is often unrealistic to assume that the behavioral policy π_t satisfies sequential ignorability (Definition 1). To address such challenges, we relax sequential ignorability of the behavioral policy, and instead posit a model of bounded confounding. We develop worst-case bounds on the evaluation policy performance $\mathbb{E}[Y(\bar{A}_{1:T})]$. In addition to the observed state $S_t(A_1^{t-1})$ available in the data, we assume that there is an *unobserved confounder* U_t available only to the behavioral policy at each time t. The behavioral policy observes the history H_t and the unobserved confounder U_t, and generates an action $A_t \sim \pi_t(\cdot \mid H_t, U_t)$. If U_t contains information about unseen potential outcomes, then sequential ignorability (Definition 1) will fail to hold for the behavioral policy.

Without loss of generality, let U_t be such that the potential outcomes are independent of A_t when controlling for U_t alongside the observed states. Such an unobserved confounder always exists since we can define U_t to be the tuple of all unseen potential outcomes.

Assumption B. *For all $t = 1, .., T$, there exists a random vector U_t such that conditional on the history H_t generated by the behavioral policy* and *U_t, $A_t \sim \pi_t(\cdot \mid H_t)$ is independent of the potential outcomes $R_t(a_{1:t}), S_{t+1}(a_{1:t}), R_{t+1}(a_{1:t+1}), S_{t+2}(a_{1:t+1}), ..,$ $S_T(a_{1:t-1}), R_T(a_{1:T})$ for all $a_{1:T} \in \mathcal{A}_1 \times \cdots \mathcal{A}_T$.*

Identification of $\mathbb{E}[Y(\bar{A}_{1:T})]$ is impossible under arbitrary unobserved confounding. However, it is often plausible to posit that the unobserved confounder U_t has a limited influence on the decisions of the behavioral policy. When the influence of unobserved confounding on each action is limited, we might expect that estimates of the evaluation policy performance assuming sequential ignorability might not be too biased. We consider the following model of unobserved confounding that bounds the influence of unobserved confounding on the behavioral policy's decisions.

Assumption C. *For $t = 1, .., T$, there is a $\Gamma_t \geq 1$ satisfying*

$$\frac{\pi_t(a_t \mid H_t, U_t = u_t)}{\pi_t(a_t' \mid H_t, U_t = u_t)} \frac{\pi_t(a_t' \mid H_t, U_t = u_t')}{\pi_t(a_t \mid H_t, U_t = u_t')} \leq \Gamma_t \tag{2}$$

for any $a_t, a_t' \in \mathcal{A}_t$, almost surely over H_t, and u_t, u_t', and sequential ignorability holds conditional on H_t ***and*** *U_t.*

When the action space is binary $\mathcal{A}_t = \{0, 1\}$, the above bounded unobserved confounding assumption is equivalent (Rosenbaum, 2002) to the following logistic model

$\log \frac{\mathbb{P}(A_t=1|H_t,U_t)}{\mathbb{P}(A_t=0|H_t,U_t)} = \kappa(H_t) + (\log \Gamma_t) \cdot b(U_t)$ for some measurable function $\kappa(\cdot)$ and a bounded measurable function $b(\cdot)$ taking values in $[0,1]$. When $T=1$, the bounded unobserved confounding assumption (2) reduces to a classical model that has been extensively studied by many authors mentioned in the related works.

Under this model of confounding, OPE is almost always unreliable; in sequential decision making, effects of confounding can create exponentially large (in the horizon T) over-sampling of large (or small) rewards, introducing an extremely large, un-correctable bias. As an illustration, consider applying OPE in the following dramatically simplified setting. Letting $U \sim \mathrm{Unif}(\{-1,1\})$ be a single unobserved confounder, consider a sequence of actions $A_1, \ldots, A_T \in \{0,1\}$ each drawn conditionally on U, but independent of one another, with the conditional distribution $P(A_t = 1 \mid U = 1) = \sqrt{\Gamma}/(1+\sqrt{\Gamma})$ and $P(A_t = 1 \mid U = 0) = 1/(1+\sqrt{\Gamma})$. Finally, consider the reward $R = U$. This reward is independent of the actions taken, yet, in the observed data, the likelihood of observing the data $((A_t = 1)_{t=1}^T, R = 1)$ is $\Gamma^{T/2}/(2(1+\Gamma)^{T/2})$, whereas the likelihood of observing $((A_t = 1)_{t=1}^T, R = 0)$ is $1/(2(1+\Gamma)^{T/2})$. Therefore, even as $n \to \infty$, OPE will mistakenly suggest that the policy which always takes $A_t = 1$ leads to much better rewards than one which always takes $A_t = 0$, because without observing U, the importance weights of both of the above samples will be equal in OPE. While this example is unrealistic, in terms of lacking states and rewards that depend on the states and actions, the core issue in terms of confounding remains: the unobserved confounder will make certain observed data samples exponentially more likely than others, without the OPE algorithm being able to tell or correct for these differences.

5. Confounding in a single decision

In many important applications, it is realistic to assume there is only a single step of confounding at a known time step t^*. Under this assumption, we outline in this section how we obtain a computationally and statistically feasible procedure for computing a lower (or upper) bound on the value $\mathbb{E}[Y(\bar{A}_{1:T})]$ of an evaluation policy $\bar{pi}$. After introducing precisely our model of confounding, we show in Proposition 1 how the evaluation policy value can be expressed using standard importance sampling weighting over steps prior to confounding time step t^* along with likelihood ratios over potential outcomes that can be used to relate the potential outcomes over observed (factual) actions with counterfactual actions not taken. These likelihood ratios over potential outcomes are unobserved, but a lower bound on the evaluation policy value can be computed by minimizing over all feasible likelihood ratios that satisfy our confounding model assumptions. Towards computational tractability, we derive a dual relaxation that can be represented as a loss minimization procedure. All proofs of results in this section are in the appendix.

We now define the confounding model for when there is an unobserved confounding variable U that only affects the behavioral policy's action at a single time period $t^\star \in [T]$. For example, in looking at impacts of confounders on antibiotics in sepsis management (Section 1.1), it is plausible to assume that after the decision when the patient arrives, unobserved confounders no longer affect later treatment decisions.

Assumption D. *For all $t \neq t^\star$, conditional on the history H_t generated by the behavioral policy, $A_t \sim \pi_t(\cdot \mid H_t)$ is independent of the potential outcomes $R_t(a_{1:t}), S_{t+1}(a_{1:t}), R_{t+1}(a_{1:t+1}), S_{t+2}(a_{1:t+1}), .., S_T(a_{1:t-1}), R_T(a_{1:T})$ for all $a_{1:T} \in \mathcal{A}_1 \times \cdots \mathcal{A}_T$. For $t = t^\star$, there exists a random variable U such that the same conditional independence holds only when conditional on the history H_t* ***and*** *U.*

We assume the unobserved confounder has bounded influence on the behavioral policy's choice of action $A_{t^\star}$:

Assumption E. *There is a $\Gamma \geq 1$ satisfying*

$$\frac{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star}, U = u)}{\pi_{t^\star}(a'_{t^\star} \mid H_{t^\star}, U = u)} \frac{\pi_{t^\star}(a'_{t^\star} \mid H_{t^\star}, U = u')}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star}, U = u')} \leq \Gamma \tag{3}$$

for any $a_{t^\star}, a'_{t^\star} \in \mathcal{A}_{t^\star}$, almost surely over $H_{t^\star}$, and u, u'.

Selecting the amount of unobserved confounding Γ is a modeling task, and the above confounding model's simplicity and interpretability makes it advantageous for enabling modelers to choose a plausible value of Γ. As in any applied modeling problem, the amount of unobserved confounding Γ should be chosen with expert knowledge (e.g. by consulting doctors that make behavioral decisions). In Section 6, we give various application contexts in which a realistic range of Γ can be posited. One of the most interpretable ways to assess the level of robustness to confounding is via the *design sensitivity* of the analysis (Rosenbaum, 2010): the value of Γ at which the bounds on the evaluation policy's value crosses a landmark threshold (e.g. performance of behavioral policy or some known safety threshold).

Under Assumption E, the likelihood ratio between the observed and unobserved distribution at $t^\star$ can at most vary by a factor of Γ. Recall that $W(a_{1:T})$ is the tuple of all potential outcomes associated with the actions $a_{1:T}$. The following observation is due to Yadlowsky et al. (2018, Lemma 2.1).

Lemma 2. *Under Assumptions D, E, for all $a_{t^\star} \neq a'_{t^\star}$, the likelihood ratio over $\{W(a_{1:T})\}_{a_{1:T}}$ exists*

$$\mathcal{L}(\cdot; H_{t^\star}, a_{t^\star}, a'_{t^\star}) := \frac{dP_W(\cdot \mid H_{t^\star}, A_{t^\star} = a'_{t^\star})}{dP_W(\cdot \mid H_{t^\star}, A_{t^\star} = a_{t^\star})},$$

and for $\mathbb{P}_W(\cdot \mid H_{t^\star}, A_{t^\star} = a_{t^\star})$-a.s. all w, w'

$$\mathcal{L}(w; H_{t^\star}, a_{t^\star}, a'_{t^\star}) \leq \Gamma \mathcal{L}(w'; H_{t^\star}, a_{t^\star}, a'_{t^\star}). \tag{4}$$

We let $\mathcal{L}(\cdot; H_{t^\star}, a_{t^\star}, a_{t^\star}) \equiv 1$. Using these (unknown) likelihood ratios, we have the following representation of $\mathbb{E}[Y(\bar{A}_{1:T})]$ under confounding.

Proposition 1. *Under Assumptions A, D, E,*

$$\mathbb{E}[Y(\bar{A}_{1:T})] = \mathbb{E}\Bigg[\prod_{t=1}^{t^\star-1}\rho_t \sum_{a_{t^\star}, a'_{t^\star}} \bar{\pi}_{t^\star}(a_{t^\star} \mid \bar{H}_{t^\star}(A_{1:t^\star}))\pi_{t^\star}(a_{t^\star} \mid H_{t^\star})$$
$$\times \mathbb{E}\Big[\mathcal{L}(W; H_{t^\star}, a_{t^\star}, a'_{t^\star}) Y_{t^\star}(a_{t^\star}) \prod_{t=t^\star+1}^{T} \rho_t \;\Big|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\Big]\Bigg],$$

using the shorthand $Y_{t^\star}(a_{t^\star}) := Y(A_{1:t^\star-1}, a_{t^\star}, A_{t^\star+1:T})$.

Proposition 1 implies a natural bound on the value $\mathbb{E}[Y(\bar{A}_{1:T})]$ under bounded unobserved confounding. Since the likelihood ratios $\mathcal{L}(\cdot; \cdot, a_{t^\star}, a'_{t^\star})$ are fundamentally unobservable due to their counterfactual nature, we take a worst-case approach over all likelihood ratios that satisfy condition (4), and derive a bound that only depend on observable distributions. Towards this goal, define the set

$$\begin{aligned}\mathfrak{L} := \big\{ & L : \mathcal{W} \times \mathcal{H}_{t^\star} \to \mathbb{R}_+ \mid L(w; H_{t^\star}) \le \Gamma L(w'; H_{t^\star}) \\ & \text{a.s. all } w, w', \text{ and } \mathbb{E}[L(W; H_{t^\star}) \mid H_{t^\star}, A_{t^\star} = a_{t^\star}] = 1\big\}. \end{aligned} \tag{5}$$

Taking the infimum over the inner expectation in the expression derived in Proposition 1, and noting that it does not depend on $a'_{t^\star}$, define

$$\eta^\star(H_{t^\star}; a_{t^\star}) := \inf_{L \in \mathfrak{L}} \mathbb{E}\left[\mathcal{L}(W; H_{t^\star}) Y_{t^\star}(a_{t^\star}) \prod_{t=t^\star+1}^{T} \rho_t \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\right].$$

Since $\eta^\star(\cdot; \cdot)$ is difficult to compute, we use functional convex duality to derive a dual relaxation that can be computed by solving a *loss minimization* problem over any well-specified model class. This allows us to compute a meaningful lower bound to $\mathbb{E}[Y(\bar{A}_{1:T})]$ even when rewards and states are continuous, by simply fitting a model using standard supervised learning methods. For $(s)_+ = \max(s, 0)$ and $(s)_- = -\min(s, 0)$, define the weighted squared loss $\ell_\Gamma(z) := \Gamma(z)_-^2 + (z)_+^2$.

Theorem 2. *Under Assumptions A, D, E, $\eta^\star(H_{t^\star}; a_{t^\star})$ is lower bounded a.s. by the unique solution $\kappa^\star(H_{t^\star}; a_{t^\star})$ to*

$$\min_{f(H_{t^\star})} \mathbb{E}\left[\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma\left(Y_{t^\star}(a_{t^\star}) \prod_{t=t^\star+1}^{T} \rho_t - f(H_{t^\star})\right)\right].$$

From Theorem 2 and Proposition 1, our final lower bound on $\mathbb{E}[Y(\bar{A}_{1:T})]$ is given by

$$\begin{aligned}&\mathbb{E}\Bigg[\prod_{t=1}^{t^\star-1}\rho_t \sum_{a_{t^\star}} \bar{\pi}_{t^\star}(a_{t^\star} \mid \bar{H}_{t^\star}(A_{1:t^\star-1})) \\ &\qquad \times (1 - \pi_{t^\star}(a_{t^\star} \mid H_{t^\star}))\kappa^\star(H_{t^\star}; a_{t^\star})\Bigg] \\ &+ \mathbb{E}\left[\pi_{t^\star}(A_{t^\star} \mid H_{t^\star}) Y(A_{1:T}) \prod_{t=1}^{T} \rho_t\right]. \end{aligned} \tag{6}$$

Note that this results in a loss minimization problem for each possible action, for each observed history H_{t^*} in the dataset generated from the behavioral policy. If confounding occurs very late in a decision process sequence, the space of histories can be very large and this may incur a significant computational cost. However if confounding occurs early in the process, the space of possible histories is small and computationally this is very tractable. This is the scenario for the domains we consider in our experiments.

Consistency We now show that an empirical approximation to our loss minimization problem yields a consistent estimate of $\kappa^\star(\cdot)$. We require the following standard overlap assumption, which states the behavioral policy has a uniformly positive probability of playing any action.

Assumption F. *There exists $C < \infty$ such that for all t and a_t, $\bar{\pi}_t(a_t \mid H_t)/\pi_t(a_t \mid H_t) \le C$ almost surely.*

Since it is not feasible to optimize over the class of all functions $f(H_{t^\star})$, we consider a parameterization $f_\theta(H_{t^\star})$ where $\theta \in \mathbb{R}^d$. We provide prove-able guarantees in the simplified setting where $\theta \mapsto f_\theta$ is linear, so that the loss minimization problem is convex. That is, we assume that f_θ is represented by a finite linear combination of some arbitrary basis functions of $H_{t^\star}$. As long as the parameterization is well-specified so that $\kappa^\star(H_{t^\star}; a_{t^\star}) = f_{\theta^\star}(H_{t^\star})$ for some $\theta^\star \in \Theta$, an empirical plug-in solution converges to $\kappa^\star$ as the number of samples n grows to infinity. We let $\Theta \subseteq \mathbb{R}^d$ be our model space; our theorem allows $\Theta = \mathbb{R}^d$.

In the below result, let $\widehat{\pi}_t(a_t \mid H_t)$ be a consistent estimator of $\pi_t(a_t \mid H_t)$ trained on a separate dataset $\mathcal{D}_n$ with the same underlying distribution; such estimators can be trained using sample splitting and standard supervised learning methods. Define the set S_ϵ of ϵ-approximate optimizers of the empirical plug-in problem

$$\min_{f(H_{t^\star})} \widehat{\mathbb{E}}_n\left[\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\widehat{\pi}_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma\left(Y_{t^\star}(a_{t^\star}) \prod_{t=t^\star+1}^{T} \widehat{\rho}_t - f(H_{t^\star})\right)\right],$$

where $\widehat{\mathbb{E}}_n$ is the empirical distribution on the data statistically independent from $\mathcal{D}_n$, and $\widehat{\rho}_t := \frac{\bar{\pi}(A_t \mid \bar{H}_t(A_{1:t-1}))}{\widehat{\pi}_t(A_t \mid H_t)}$.

We assume that we observe independent, and identically distributed trajectories, and formally, assume that the observed

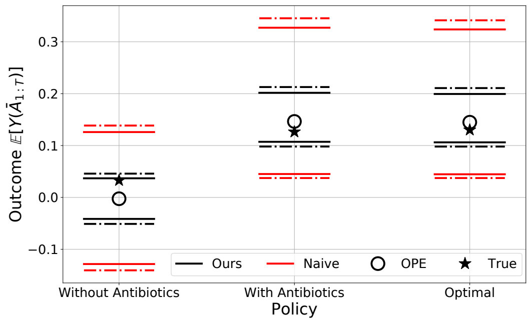
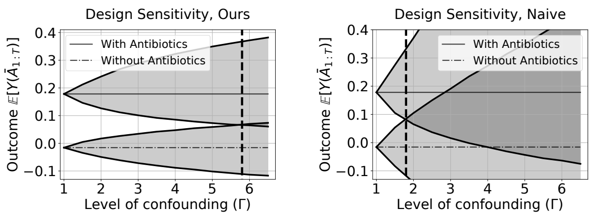


Figure 1. Sepsis simulation. Data generation process with the level of confounding $\Gamma^\star = 2.0$. Estimated outcome with OPE along with the true value. Black lines show estimated upper and lower bound on outcome using our approach and red lines correspond to the naïve approach, both with $\Gamma = 2.0$. Dashed lines represents 95% quantile.



(a) Our approach (b) Naive approach

Figure 2. Sepsis simulator design sensitivity. Data generation process with level of confounding $\Gamma^\star = 5$. Estimated lower and upper bound of two policies (with and without antibiotics) under (a) our approach with sensitivity 5.8 (b) naive approach with sensitivity 1.8.

cumulative reward is the evaluation of the potential outcome at the observed action sequence, $Y_{\text{obs}} = Y(A_{1:T})$ so that each trajectory (unit) does not affect one another[1].

Theorem 3. *Let Assumptions A, D, E, F hold, and let $\theta \mapsto f_\theta$ be linear such that $f_{\theta^\star}(\cdot) = \kappa^\star(\cdot, a_{t^\star})$ for some unique $\theta^\star \in \mathbb{R}^d$. Let $\mathbb{E}|Y_{t^\star}(a_{t^\star})|^4 < \infty$, and $\mathbb{E}[|f_\theta(H_{t^\star})|^4] < \infty$ for all $\theta \in \Theta$. If for all t, $\widehat{\pi}_t(\cdot|\cdot) \to \pi_t(\cdot|\cdot)$ pointwise a.s., $\bar{\pi}_t(\cdot|\cdot)/\widehat{\pi}_t(\cdot|\cdot) \le 2C$, and $\exists c$ s.t. $0 < c \le \widehat{\pi}_{t^\star}(a_{t^\star}|H_{t^\star}) \le 1$ a.s., then $\liminf_{n\to\infty} dist(\theta^\star, S_{\varepsilon_n}) \xrightarrow{p} 0\ \forall \varepsilon_n \downarrow 0$.*

Hence, a plug-in estimator of the lower bound (6) is consistent as $n \to \infty$, under the hypothesis of Theorem 3.

6. Experiments

We provide a number of examples of how our method could be applied in real off-policy evaluation settings, where confounding is primarily an issue in a single decision within the sequence. After introducing these settings and why our model for confounding might fit these settings, we demonstrate using the method to certify the reliability of (or raise concerns about the unreliability of) OPE in these settings. Because the gold standard real counterfactual outcomes are only known in simulations, we focus on simulation examples motivated by the real OPE applications. First, we introduce the real world setting, the corresponding simulators, and how we introduce confounding in the simulator to model the realistic source of confounding that might exist in off-policy data in these settings. Then, we use these examples to demonstrate that our approach can be fairly tight in some cases, meaning that our bounds are close to the true evaluation policy performance after introducing confounding in our simulations and applying our method. We also demonstrate how our method compares to the naïve approach in allowing us to certify robustness to confounding with much larger values of Γ than the naïve approach.

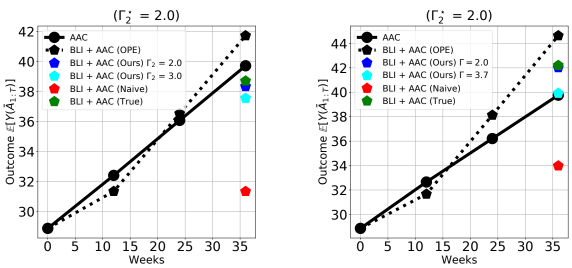
Managing sepsis patients To simulate data as in the example in Section 1.1, we used the sepsis simulator developed by Oberst and Sontag (2019). To simulate the unrecorded co-morbidities that introduce confounding, we extract some of the randomness that goes into choosing the state transitions into a confounding variable, so that the confounding variable are correlated with better state transitions in the simulation. In the first time step, we take the optimal action with respect to all other drugs, and select antibiotics with probability $\sqrt{\Gamma}/(1+\sqrt{\Gamma})$ if the confounding variable is large and with probability $1/(1+\sqrt{\Gamma})$ if the confounding variable is small, satisfying Assumption E. We assume that the care team acts nearly optimally, except for some randomness due to the challenges of the ICU, guaranteeing overlap (Assumption F) with respect to the optimal evaluation policy. In all but the first time step, we implemented the behavior policy to take the optimal next treatment action with probability 0.85, and otherwise switch the vasopressor status, independent of the confounders, satisfying Assumption D.

We imagine that using existing medical knowledge, an automated policy is implemented to implement optimal treatment policy, and we would like to evaluate it's benefit relative to the current standard of care. We learn optimal policy with respect to this simulation online (without confounding) using policy iteration, as done in Oberst and Sontag (2019).

Communication interventions for minimally verbal children with autism Minimially verbal children represent 25-30% of children with autism, and often have poor prognosis in terms of social functioning (Rutter et al., 1967; Anderson et al., 2009). See Kasari et al. (2014) for more background on the challenges of treating these patients.

We compare the number of speech utterances by such children under an adaptive policy that starts with behavioral language interventions (BLI) for 12 weeks and augments BLI with an augmented or alternative communication (AAC) approach against a non-adaptive policy that uses AAC through the whole treatment. Kasari et al. (2014) note that there are very few randomized trials of these interventions, and the

[1] together, these imply the stable unit treatment (SUTVA) assumption (Rubin, 1980) in the statistics literature



(a) Case I (b) Case II

Figure 3. Autism simulation. Outcome of two different policies, confounded adaptive policy (BLI+AAC) and unconfounded non-adaptive policy (AAI). Data generation process with the level of confounding $\Gamma^{\star} = 2.0$. Case I: effect size 0.3. Case II: effect size 0.8

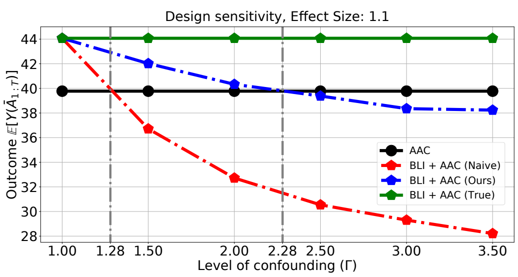


Figure 4. Autism simulation design sensitivity. Data generation process with the level of confounding $\Gamma^{\star} = 1.0$. True value of adaptive (BLI+AAC) and non-adaptive (AAC) policies along with estimated lower bound on outcome using our and naive approach

number of individuals in these trials tends to be small. It is natural in such cases to consider using existing off-policy data to evaluate this intervention protocol.

At the beginning of the treatment children are assigned to BLI or AAC treatment pesudo-randomly due to the availability of AAC devices. However, at the follow up visit after 12 weeks a clinician may decide to use AAC devices for some children starting with BLI. Since this interventions requires a specialized device, it is likely that the clinicians working with the children only give the devices to those for whom the device is most effective. Assessing the effectiveness of the intervention is likely based on the clinician's interactions with the patients, not information encoded in the reported covariates, which contain partial, noisy information about the outcome. Therefore, while there is confounding, Assumption D is plausible in the second decision.

The simulation for comparing developmental interventions for autistic children comes from Lu et al. (2016) based on modeling the data from Kasari et al. (2014). Lu et al. (2016) provide plausible ranges for the parameters of the simulation, based on the observed results of the SMART trial and realistic effect sizes. We create the aformentioned confounding variable in our experiments by making the variables in the simulation that corresponds to the effectiveness of the intervention a randomly selected value that is unobserved. We introduce confounding by simulating the decision in the second time step based on this latent variable, in accordance with the model in Assumptions D and E.

Results All implementation and model details can be found in the appendix. We compare three different approaches: applying standard OPE methods that assume sequential ignorability holds, computing lower- and upper-bounds on the evaluation policy performance using the naïve bound provided in the Appendix, and computing these bounds using our proposed loss minimization approach.

Sepsis simulator We evaluate three different policies, 1. Without antibiotics (WO), which does not administer antibiotics at the first timestep, 2. With antibiotics (W), which administer antibiotics at the first time step, and 3. the optimal policy learned by policy iteration. Figure 1 shows the outcome of these policies estimated on the data generated with $\Gamma^{\star} = 2.0$. Confounding leads standard OPE methods to underestimate the outcome for WO policy and over estimate the outcome for optimal and W policy, which makes W and optimal policy looks much better than WO. The naive bound cannot guarantee the superiority of W and optimal policy over WO with $\Gamma = 2.0$; however, our proposed method shows the lower bound on W and optimal do not cross the upper bound on WO, certifying the robustness of the benefit of immediately administering antibiotics.

Figure 2 compares the design sensitivity of our method versus the naive approach. We generated the data with $\Gamma^{\star} = 5.0$. Figure 2(a,b) shows that using our method (respectively naïve), the lower bound on W policy meets the upper bound of WO policy at $\Gamma = 5.8$ (respectively $\Gamma = 1.8$). This indicates the improved robustness of our algorithm to conservative choices of Γ.

Autism simulator We consider two different cases. Case I, effect size 0.3 (Figure 3): the adaptive policy (BLI+AAC) has lower true outcome than the non-adaptive policy (AAC). We injected $\Gamma^{\star} = 2.0$ level of confounding in this simulation that makes the standard OPE approach over estimate the outcome of the adaptive policy. However, by using our method to compute a lower bound on the adaptive policy, Figure 3 (a) shows that we cannot guarantee this superiority with the amount of confounding $\Gamma = 2.0$. Case II, effect size 0.8: the adaptive policy has a higher outcome than the non adaptive policy with this effect size, and with the amount of confounding $\Gamma^{\star} = 2.0$, standard OPE methods overestimate this value. Figure 3(c) shows that unlike the naïve method, our method guarantees the superiority of the adaptive policy by $\Gamma \in [2.0, 3.7]$. Figure 4 shows the design sensitivity of our method. In this example the effect size is 1.1 and the generated data is unconfounded. Lower bound computed by the naïve method shows design sensitivity of $\Gamma = 1.28$ while using our method is robust to more conservative choices of $\Gamma = 2.28$.

References

D. K. Anderson, R. S. Oti, C. Lord, and K. Welch. Patterns of growth in adaptive social abilities among children with autism spectrum disorders. *Journal of Abnormal Child Psychology*, 37(7):1019–1034, Oct 2009. ISSN 1573-2835. doi: 10.1007/s10802-009-9326-0. URL https://doi.org/10.1007/s10802-009-9326-0.

A. J. Brent. Meta-analysis of time to antimicrobial therapy in sepsis: Confounding as well as bias. *Critical Care Medicine*, 45(2), 2017.

B. A. Brumback, M. A. Hernán, S. J. P. A. Haneuse, and J. M. Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, 23(5):749–767, 2004.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22 (1):173–203, 1959.

J. Futoma, A. Lin, M. Sendak, A. Bedoya, M. Clement, C. O'Brien, and K. Heller. Learning to treat sepsis with multi-output gaussian process deep recurrent q-networks, 2018. URL https://openreview.net/forum?id=SyxCqGbRZ.

O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi. Guidelines for reinforcement learning in healthcare. *Nat Med*, 25(1): 16–18, 2019a.

O. Gottesman, Y. Liu, S. Sussex, E. Brunskill, and F. Doshi-Velez. Combining parametric and nonparametric models for off-policy evaluation. In *International Conference on Machine Learning*, pages 2366–2375, 2019b.

J. Hanna, S. Niekum, and P. Stone. Importance sampling policy evaluation with an estimated behavior policy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, June 2019.

J. P. Hanna, P. Stone, and S. Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

M. Hernán and J. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

M. D. Howell and A. M. Davis. Management of Sepsis and Septic Shock. *JAMA*, 317(8):847–848, 02 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.0131. URL https://doi.org/10.1001/jama.2017.0131.

T.-C. Hu, F. Moricz, and R. Taylor. Strong laws of large numbers for arrays of rowwise independent random variables. *Acta Mathematica Hungarica*, 54(1-2):153–162, 1989.

G. W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2): 126–132, 2003.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*, 2019.

N. Kallus and A. Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, pages 9269–9279, 2018.

N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. *arXiv preprint arXiv:1810.02894*, 2018.

C. Kasari, A. Kaiser, K. Goods, J. Nietfeld, P. Mathy, R. Landa, S. Murphy, and D. Almirall. Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 53(6):635–646, 2014.

A. J. King and R. J. Wets. Epi-consistency of convex stochastic programs. *Stochastics and Stochastic Reports*, 34(1-2): 83–92, 1991.

M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018a.

M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018b.

H. M. Le, C. Voloshin, and Y. Yue. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.

Y. Liu, O. Gottesman, A. Raghu, M. Komorowski, A. A. Faisal, F. Doshi-Velez, and E. Brunskill. Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2644–2653, 2018.

X. Lu, I. Nahum-Shani, C. Kasari, K. G. Lynch, D. W. Oslin, W. E. Pelham, G. Fabiano, and D. Almirall. Comparing dynamic treatment regimes using repeated-measures outcomes: modeling considerations in smart studies. *Statistics in medicine*, 35(10):1595–1615, 2016.

D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

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

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

S. A. Murphy, M. J. van der Laan, and J. M. Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

X. Nie, E. Brunskill, and S. Wager. Learning when-to-treat policies. *arXiv preprint arXiv:1905.09751*, 2019.

M. Oberst and D. Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4881–4890, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/oberst19a.html.

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

A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. *arXiv preprint arXiv:1705.08422*, 2017.

A. Rhodes, L. E. Evans, W. Alhazzani, et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. *Intensive Care Medicine*, 43(3):304–377, 2017.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

J. M. Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

J. M. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. E. Halloran and D. Berry, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 1–94, New York, NY, 2000. Springer New York. ISBN 978-1-4612-1284-3.

R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, New York, 1998.

P. R. Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

P. R. Rosenbaum. *Design of Observational Studies*, volume 10. Springer, 2010.

P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

D. B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

M. Rutter, D. Greenfeld, and L. Lockyer. A five to fifteen year follow-up study of infantile psychosis: Ii. social and behavioural outcome. *British Journal of Psychiatry*, 113 (504):1183–1199, 1967. doi: 10.1192/bjp.113.504.1183.

C. W. Seymour, F. Gesten, H. C. Prescott, M. E. Friedrich, T. J. Iwashyna, G. S. Phillips, S. Lemeshow, T. Osborn, K. M. Terry, and M. M. Levy. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244, 2017. doi: 10.1056/NEJMoa1703058. URL https://doi.org/10.1056/NEJMoa1703058. PMID: 28528569.

S. A. Sterling, W. R. Miller, J. Pryor, M. A. Puskarich, and A. E. Jones. The impact of timing of antibiotics on outcomes in severe sepsis and septic shock: a systematic review and meta-analysis. *Critical care medicine*, 43(9): 1907, 2015.

P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

P. S. Thomas, B. C. da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

S. Yadlowsky, H. Namkoong, S. Basu, J. Duchi, and L. Tian. Bounds on the conditional and average treatment effect in the presence of unobserved confounders. *arXiv:1808.09521 [stat.ME]*, 2018.

J. Zhang and E. Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13401–13411. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9496-near-optimal-reinforcement-learning-in-dynamic-treatment-regimes.pdf.