Learning Models and Evaluating Policies with Offline Off-Policy Data under Partial Observability

Shreyas Chaudhari SCHAUDHARI@CS.UMASS.EDU
University of Massachusetts

Philip S. Thomas PTHOMAS@CS.UMASS.EDU
University of Massachusetts

Bruno Castro da Silva BSILVA@CS.UMASS.EDU
University of Massachusetts

Abstract

Models in reinforcement learning are often estimated from offline data, which in many real-world scenarios is subject to partial observability. In this work, we study the challenges that emerge from using models estimated from partially-observable offline data for policy evaluation. Notably, a complete definition of the models includes dependence on the data-collecting policy. To address this issue, we introduce a method for model estimation that incorporates importance weighting in the model learning process. The off-policy samples are reweighted to be reflective of their probabilities under a different policy, such that the resultant model is a consistent estimator of the off-policy model and provides consistent estimates of the expected off-policy return. This is a crucial step towards the reliable and responsible use of models learned under partial observability, particularly in scenarios where inaccurate policy evaluation can have catastrophic consequences. We empirically demonstrate the efficacy of our method and its resilience to common approximations such as weight clipping on a range of domains with diverse types of partial observability.

Keywords: Partial Observability, Model Misspecification, Off-Policy Evaluation

1. Introduction

Many recent successes of reinforcement learning have been facilitated by the use of models (Degrave et al., 2022; Mirhoseini et al., 2020; Mannion et al., 2016; Schrittwieser et al., 2020). Models provide the ability to predict the downstream effects of current decisions, which is useful in many problems such as control, planning, aiding directed exploration and performing policy evaluation. However, it is uncommon to have access to a perfect model of the system under consideration. Most real-world scenarios require learning models from previously collected data—often called *offline* data. This approach is particularly prevalent in safety-critical settings, where online data collection may be risky or infeasible (Dulac-Arnold et al., 2021). In this paper, we address an important challenge arising from the need to learn models from data: *in most real-world problems, the data captured about a given system seldom includes all relevant information about the state of the system.* Consider, for example, a medical application where it may not be feasible to measure all attributes of a patient—such as their genetic profile or exposure to pollution—despite these attributes being potentially relevant in designing appropriate treatments.

We explore the challenges that emerge due to this *partial observability* for using models learned from partially-observable offline data, particularly for policy evaluation. Standard methods for model estimation developed for the fully-observable case when applied to this setting may result in arbitrarily inaccurate policy evaluation. We first demonstrate that models constructed from partially-observable data depend strongly on the policy employed to gather training data. Crucially, these models cannot be used to reliably evaluate other policies, as they may yield arbitrarily incorrect return estimates.[1] Addressing this challenge is important since policy evaluation techniques are key building blocks for a wide range of tasks in reinforcement learning.

In this paper, we introduce a method that uses offline data with partial state information (called *observations*) to construct models that are guaranteed to produce consistent off-policy estimates of the performance of policies of interest. We refer to models over observations as *partial models*. Unlike models learned in the fully-observable setting, partial models are not policy-invariant: even if trained with infinite data, the estimated models would in general be a function of the policy used to gather training data (Figure 1). To address this policy-dependence and make the partial models amenable for off-policy evaluation, we propose a novel technique for model estimation that integrates importance sampling (IS) into the model learning process. The approach uses importance weights to reweight the frequency of the samples in the offline data such that they are reflective of the expected frequencies with which they would have occurred under the policy being evaluated—rather than the data-collection policy. The models so learned from off-policy partially observable data are consistent estimates of the off-policy models (Theorem 3) and are guaranteed to produce consistent estimates of the off-policy expected return (Theorem 4). We validate our theoretical results with numerical experiments and empirically demonstrate the efficacy of our method and its resilience to standard approximations like weight-clipping on a range of domains with different types of partial observability.

2. Problem Setup

Problem Setup: Our goal is to derive models from partially-observable offline data that offer theoretical guarantees for the task of policy evaluation. Specifically, these models should be guaranteed to yield consistent off-policy return estimates. We model this setting as an episodic partially-observable Markov decision process (POMDP) (Lovejoy, 1991). A POMDP is tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, p, \Omega, R, d_0)$, where $\mathcal{S}$ is a set of states, $\mathcal{O}$ is a set of discrete observations, $\mathcal{A}$ is a set of discrete actions, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function, $\Omega : \mathcal{S} \to \mathcal{O}$ is the observation function, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, and $d_0 : \mathcal{S} \to [0, 1]$ is the initial state distribution. The underlying unobserved Markov decision process M is denoted by $(\mathcal{S}, \mathcal{A}, p, R, d_0)$. We assume that the observation function Ω is a many-to-one mapping from states to observations. This aligns with the more general setting where states cannot be directly inferred from observations.

An offline dataset is constructed by deploying some data-collecting policy, $\pi_b : \mathcal{O} \times \mathcal{A} \to [0, 1]$, also referred to as the *behavior policy.*[2] The data set consists of n observed trajectories, $\mathcal{D}_n^{(\pi_b)} \triangleq (H_i)_{i=1}^n$, where each H_i is a trajectory $(O_0, A_0, R_0, O_1, \ldots)$. Notice

[1] This phenomenon has been studied in the econometrics literature, where it is known as the Lucas Critique (Brunner et al., 1983), as well as in recent works in reinforcement learning (Rezende et al., 2020).

[2] All our analyses remain valid in scenarios where data is gathered from multiple behavioral policies.

that trajectories H_i do not contain the underlying states, S_t, for any time steps t. We denote the model estimated from the offline dataset, $\mathcal{D}_n^{(\pi_b)}$, as $\hat{M}_n^{(\pi_b)} \triangleq (\mathcal{O}, \mathcal{A}, \hat{p}_n, \hat{R}_n, \hat{d}_{0,n})$, where the superscript denotes the policy used for collecting the data used to estimate such a model. Here, $\hat{M}_n^{(\pi_b)}$ is modeled as a Markov decision process (MDP) over observations, where $\hat{p}_n : \mathcal{O} \times \mathcal{A} \times \mathcal{O} \to [0, 1]$, $\hat{R}_n : \mathcal{O} \times \mathcal{A} \times \mathcal{O} \to \mathbb{R}$, and $\hat{d}_{0,n} : \mathcal{O} \to [0, 1]$ are the corresponding maximum likelihood estimates given by:

$$\hat{d}_{n,0}(o) := \frac{\sum_n \mathbf{1}_0^i\{o\}}{n}; \; \hat{p}_n(o, a, o') := \frac{\sum_{n,t} \mathbf{1}_t^i\{o, a, o'\}}{\sum_{n,t} \mathbf{1}_t^i\{o, a\}}; \; \hat{R}_n(o, a, o') := \frac{\sum_{n,t} \mathbf{1}_t^i\{o, a, o'\} R_t^i}{\sum_{n,t} \mathbf{1}_t^i\{o, a, o'\}} \tag{1}$$

This estimated model is called a **partial** model since it models transitions between partially observable states (observations), given an action, along with a reward function that maps observations to scalars. To perform off-policy evaluation (OPE) with these partial models, we make the support assumption standard in the OPE literature.

Assumption 1 $\exists\, \varepsilon > 0 : (\pi_b(o, a) < \varepsilon) \implies (\pi_e(o, a) = 0)$, *for all* $o \in \mathcal{O}$ *and* $a \in \mathcal{A}$.

Notation: The indicator function is denoted by $\mathbf{1}\{\cdot\}$, and $\mathbf{1}_t\{\cdot\}$ is the indicator function for a condition at time t. For example, $\mathbf{1}_t\{o, a, o'\} \triangleq \mathbf{1}\{O_t = o, A_t = a, O_{t+1} = o'\}$. Asymptotic convergence targets of finite-sample estimates when $n \to \infty$ are denoted using the subscript ∞. For example, $\hat{M}_\infty \triangleq (\hat{p}_\infty, \hat{R}_\infty, \hat{d}_{0,\infty})$ denotes $\lim_{n\to\infty} \hat{M}_n$. The expected return of a policy π, deployed on environment M, is denoted by $J(\pi; M) = \mathbb{E}_\pi \left[\sum_t R_t\right]$.

3. Challenges Arising from Partial Observability

Estimating models from offline data in the *fully observable* setting has been widely studied in reinforcement learning (Janner et al., 2019). The primary challenges in that setting are those of partial coverage (Kidambi et al., 2020; Uehara and Sun, 2021) and the need for a sufficiently-expressive parametric model class (Lu et al., 2021). The presence of partial observability in the dataset introduces challenges orthogonal to the aforementioned ones, hindering the direct application of existing methods. The challenges, specifically, are 1) the policy-dependence of the estimated models, and 2) the subsequent inaccuracy in off-policy evaluation. We elucidate both with illustrative examples.

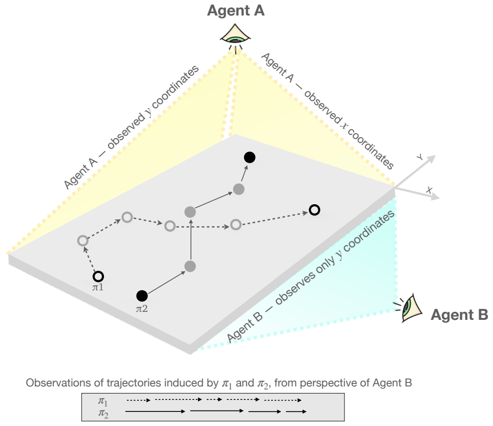
Agent A
Agent A — observed y coordinates
Agent A — observed x coordinates
Y
X
π1
π2
Agent B — observes only y coordinates
Agent B
Observations of trajectories induced by π_1 and π_2, from perspective of Agent B
π_1
π_2

Figure 1: Policy-dependent partial models.

Policy-Dependence: The partial models (Equation (1)) must be defined in conjunction with the data-collection policy, as they are not invariant to changing data-collecting policies. For instance, consider the MDP in Figure 1, in which a point mass traverses a 2D plane. Two potential trajectories, resulting from deploying different policies (π_1 and π_2) are shown. Agent-A can observe all state features (x and y coordinates of the point mass), while Agent-B, can only perceive the y coordinates of each trajectory. An offline dataset collected by Agent-A, thus contains information about complete Markov states of the system, and can be used to learn the true underlying model of the MDP (assume full coverage and an expressive model class). On the contrary, in the dataset collected by Agent-B, the same action from the same observation may result in different outcomes, i.e., the next observation and reward, *depending on the data-collecting policy* used. Thus the partial model estimated by will be different for different data-collecting policies, and must be defined in conjunction with the policy deployed to collect the offline data.

Inaccurate Off-Policy Evaluation: The partial models cannot be used to accurately evaluate policies other than the data-collecting policy. Consider the MDP depicted in Figure 2, where the five states are represented by circles and four observations by boxes. The observation o_2 aliases states s_2 and s_3 into a single observation. The partial reward functions estimated by the two data-collecting policies, $\pi_1 = [a_u, a_u]$ and $\pi_2 = [a_d, a_u]$, differ due to policy-dependence of the models. Specifically, for the input (o_2, a_u, o_3) the reward esimated by π_1 would be $\hat{R}^{(\pi_1)}(o_2, a_u, o_3) = +10$, whereas with π_2 it would be $\hat{R}^{(\pi_2)}(o_2, a_u, o_3) = -10$. Consequently, using the partial model estimated using π_1 as the data-collecting policy for evaluating π_2 would result in incorrect return estimates.

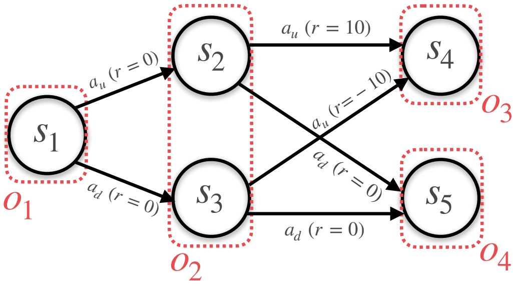

Figure 2: Inaccurate off-policy evaluation due to policy-dependence of partial models.

4. Our Approach: Importance-Weighted Partial Models

The previous section highlights the challenges posed by partial observability on model learning and policy evaluation using partial models. Despite the model estimates from different data-collecting policies differing, the respective models surprisingly *can* be used to obtain consistent and accurate estimates of the expected return for the *data-collecting policy* itself.

Lemma 2

The result implies that in theory a model estimated from data collected by the *evaluation policy* can be used for off-policy evaluation. Since the underlying process (M) governing both partial models is the same, the model parameters estimated from the two policies only differ due to differences in state visitation distribution (Equations (7) and (8)) induced by a different policy. We show that the effects of the change in state visitation distribution can be addressed using **importance sampling** (Horvitz and Thompson, 1952; Precup, 2000; Ripley, 2009). To estimate a partial model as seen under the evaluation policy, we want the *probabilities* (or, visitation counts of samples) of terms in Equation (1) to be reflective of those seen when π_e is executed. This can be attained by multiplying the indicator terms by an importance weight ρ_t in Equation (1) to estimate model components.

$$\hat{p}_n^{IWPM}(o, a, o') := \frac{\sum_{n,t} \mathbf{1}_t^i\{o, a, o'\}\rho_t^i}{\sum_{n,t} \mathbf{1}_t^i\{o, a\}\rho_t^i}; \qquad \hat{R}_n^{IWPM}(o, a, o') := \frac{\sum_{n,t} \mathbf{1}_t^i\{o, a, o'\}R_t^i\rho_t^i}{\sum_{n,t} \mathbf{1}_t^i\{o, a, o'\}\rho_t^i} \tag{2}$$

where, $\rho_t^i = \frac{\Pr(H_t^i;\pi_e)}{\Pr(H_t^i;\pi_b)} = \prod_{j=1}^t \frac{\pi_e(O_j^i, A_j^i)}{\pi_b(O_j^i, A_j^i)}$. This method for estimating importance-weighted (off-policy) partial models is abbreviated as IWPM . The model estimates in

Equation (2) asymptotically converge to the partial model induced by π_e (Theorem 3) and can thus be used for consistent off-policy return estimation (Theorem 4).

Theorem 3 *The importance-weighted partial model* $\hat{M}_n^{(\pi_e;IWPM)}$ *is a consistent estimator of the model* $\hat{M}_\infty^{(\pi_e)}$ *estimated from* $D_\infty^{(\pi_e)}$*, i.e.,* $\hat{M}_n^{(\pi_e;IWPM)} \xrightarrow{a.s.} \hat{M}_\infty^{(\pi_e)}$.

Theorem 4 *Using the importance-weighted partial model* $\hat{M}_n^{(\pi_e;IWPM)}$ *to evaluate* π_e *is guaranteed to provide consistent estimates of its true expected return, i.e.,* $J(\pi_e; \hat{M}_n^{(\pi_e;IWPM)}) \xrightarrow{a.s.} J(\pi_e; M)$.

The proofs for all results may be found in the Appendix. To our knowledge, this is the first method that addresses the problem of model-based off-policy evaluation using offline partially observable data without any additional assumptions.

5. Empirical Evaluation

First, we conduct diagnostic experiments providing intuition for our theoretical results and an understanding of the properties of our method. Then we proceed to evaluate the effectiveness of importance-weighted partial models for off-policy evaluation across domains with diverse kinds of partial observability. Experimental details, along with descriptions of the domains, can be found in the Appendix.

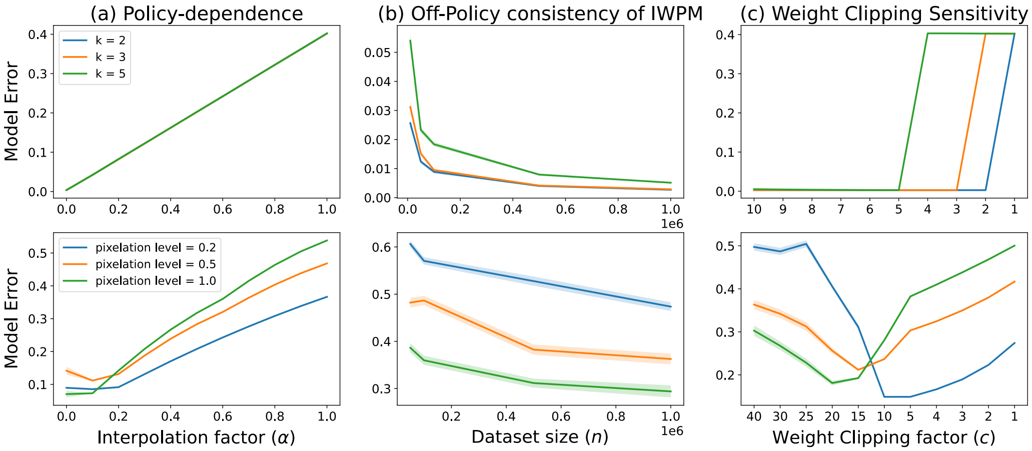

Figure 3: Diagnostic experiments on (top) KMarkovGraph and (bottom) PointEnv.

Diagnostic Experiments: These are conducted on two domains: a k-th order POMDP structured graph, KMarkovGraph, and a 2D point navigation environment, PointEnv.
(a) *Policy-dependence of partial models:* The maximum likelihood model $\hat{p}_n^{(\pi)}$ (Equation (1)) is estimated from offline datasets collected by executing different policies $\pi^{(\alpha)} = \alpha\pi_b + (1-\alpha)\pi_e$ that interpolate between π_b and π_e, with $\alpha \in [0, 1]$. The model mismatch error—measured by the total variation distance—between $\hat{p}_n^{(\pi^{(\alpha)})}$ and $\hat{p}_n^{(\pi_e)}$, the ground truth model estimated from off-policy data $\mathcal{D}_n^{(\pi_e)}$, as a function of α showcases that the policy dependence in Figure 3 (a). As the policy increasingly deviates from π_e with increasing values of α, the model mismatch error also increases.

(b) *Consistency of the importance weighted partial model:* The size (n) of the offline dataset $\mathcal{D}_n^{(\pi_b)}$ used for estimating the importance-weighted partial model $\hat{p}^{(\pi_e;\text{IWPM})}$ is varied. It is observed that model error decreases with increasing n, as expected.

(c) *Effect of weight clipping:* To test the sensitivity of model estimation to weight clipping, we vary the weight clipping factor c used for computing importance weights. For smaller values of c, we would expect to see higher bias, but lower variance in estimation and the opposite for larger values as is seen in the plot for PointEnv. Due to the k-th order Markov nature of KMarkovGraph, values of $c < k$ incur model error in estimation while values greater than c result in negligible model error.

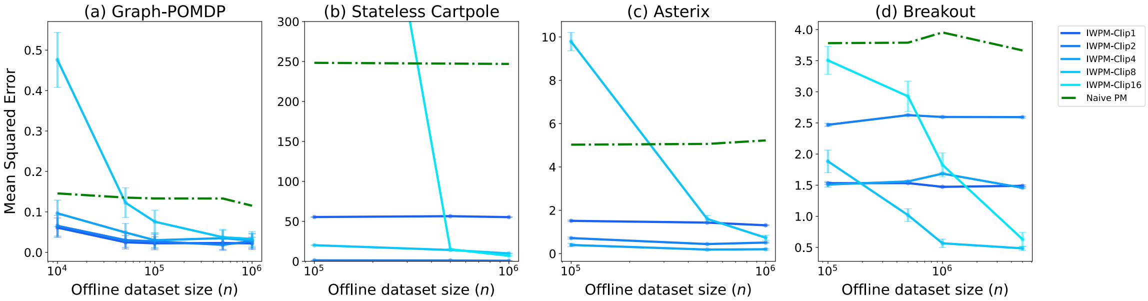

Figure 4: Mean-squared-error for OPE using IWPM, for varying n.

Off-Policy Evaluation: To contextualize our results we compare against naively estimating partial models that are maximum likelihood model estimates from $\mathcal{D}_n^{(\pi_b)}$, labeled Naive PM. A comparison with model-free estimators can be found in the Appendix. We evaluate IWPM with varying levels of weight clipping, labeled as IWPM-Clip-c, where c denotes the weight clipping factor. For our experiments, we use environments from various OPE benchmarks: (a) Graph-POMDP (Voloshin et al., 2021), (b) Stateless Cartpole (Morad et al., 2023), and two Atari games (c) Asterix, and (d) Breakout from the MinAtar testbed (Young and Tian, 2019). Results in Figure 4 indicate that the naive method incurs an irreducible bias in most domains since it relies on a model that is a function of π_b while evaluating π_e. For IWPM, we observe that weight clipping often aids the performance of the method due to lower variance in estimation and allows the method to scale up to long horizon problems—for instance, the Asterix domain has an average episode length of approximately 100. As indicated by the standard error bars, the method exhibits low variance in estimation. We report these results for different sizes of the offline dataset, and as expected the estimates get increasingly accurate with larger dataset size n. Thus IWPM presents a viable approach for conducting model-based off-policy evaluation under partial observability.

6. Conclusion

In this paper, we consider the problem of model estimation and off-policy evaluation from offline data with partial observability. We highlight the challenges posed by partial observability for the task that renders existing methods ineffective. We propose a model estimation procedure that incorporates importance weights to address those challenges. We empirically validate our theory and the efficacy of our method for off-policy evaluation on partially-observable domains, and see it scale to long horizons.

References

Cameron Allen, Neev Parikh, Omer Gottesman, and George Konidaris. Learning markov state abstractions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:8229–8241, 2021.

Karl Brunner, Alan Meltzer, et al. Econometric policy evaluation: A critique. In *Theory, Policy, Institutions: Papers from the Carnegie-Rochester Conferences on Public Policy*, volume 1, page 257. North Holland, 1983.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602 (7897):414–419, 2022.

Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685, 1952.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Didolkar, Dipendra Misra, Dylan Foster, Lekan Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of controllable latent states with multi-step inverse models. *arXiv preprint arXiv:2207.08229*, 2022.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. In *AI&M*, 2006.

William S Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28(1):47–65, 1991.

Cong Lu, Philip J Ball, Jack Parker-Holder, Michael A Osborne, and Stephen J Roberts. Revisiting design choices in offline model-based reinforcement learning. *arXiv preprint arXiv:2110.04135*, 2021.

Patrick Mannion, Jim Duggan, and Enda Howley. An experimental review of reinforcement learning algorithms for adaptive traffic signal control. *Autonomic road transport support systems*, pages 47–66, 2016.

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Sungmin Bae, et al. Chip placement with deep reinforcement learning. *arXiv preprint arXiv:2004.10746*, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POPGym: Benchmarking partially observable reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=chDrutUTs0K`.

Khanh Xuan Nguyen. Converting pomdps into mdps using history representation. 2021.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Danilo J Rezende, Ivo Danihelka, George Papamakarios, Nan Rosemary Ke, Ray Jiang, Theophane Weber, Karol Gregor, Hamza Merzic, Fabio Viola, Jane Wang, et al. Causally correct partial models for reinforcement learning. *arXiv preprint arXiv:2002.02836*, 2020.

Brian D Ripley. *Stochastic simulation.* John Wiley & Sons, 2009.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609, 2020.

Pranab K Sen and Julio M Singer. *Large sample methods in statistics: an introduction with applications*, volume 25. CRC press, 1994.

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-dependent value-based off-policy evaluation in pomdps. *arXiv preprint arXiv:2207.13081*, 2022.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL `https://openreview.net/forum?id=IsK8iKbL-I`.

Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.