POPCORN: Partially Observed Prediction Constrained Reinforcement Learning

Joseph Futoma
Harvard SEAS

Michael C. Hughes
Tufts University, Dept. of Computer Science

Finale Doshi-Velez
Harvard SEAS

Abstract

Many medical decision-making tasks can be framed as partially observed Markov decision processes (POMDPs). However, prevailing two-stage approaches that first learn a POMDP and then solve it often fail because the model that best fits the data may not be well suited for planning. We introduce a new optimization objective that (a) produces both high-performing policies and high-quality generative models, even when some observations are irrelevant for planning, and (b) does so in batch off-policy settings that are typical in healthcare, when only retrospective data is available. We demonstrate our approach on synthetic examples and a challenging medical decision-making problem.

1 Introduction

Reinforcement learning (RL) has the potential to assist sequential decision-making in healthcare, especially in settings lacking strong evidence-based guidelines. For example, in this work we will focus on the task of managing patients in an intensive care unit (ICU) with acute hypotension, a life-threatening emergency in which a patient's blood pressure drops dangerously low. In situations like this in critical care, it is often unclear which treatment will be most effective for a given patient, and in what amount, frequency, and duration (García et al., 2015). RL might help answer these questions, but applying RL in healthcare is challenging, as highlighted by Gottesman et al. (2019), and we are still far from integration into the clinic. Two key challenges for most clinical decision-making problems, including ours, are:

1. Medical environments are *partially observable*: a patient's current physiological state alone is insufficient to make good decisions, and we need other context about their history.
2. We must learn in a *batch* setting, given only a single batch of retrospective (usually observational) data.

In this work we focus on pushing the limits of *model-based* RL, using *discrete* hidden state representations. Generative models can of course help us learn to recommend good actions (our primary goal), but they also have many other important benefits. For instance, we can use them to predict future observations (a form of validation), they can learn in the presence of missing data (pervasive in clinical settings), and they are generally more sample-efficient than competing model-free approaches (important as many medical problems are data-limited). Building directly inspectable models via simple, discrete structures further enables easy inspection for clinical sensibility, a task much harder and sometimes impossible to accomplish from more complicated black-box models (e.g. deep learning).

We propose POPCORN[1], or Partially Observed Prediction Constrained Reinforcement learning, a new optimization objective for the well-known partially observable Markov decision process (POMDP) (Kaelbling et al., 1998). POMDPs have been traditionally trained in a two-stage process, where the first stage learns a generative model by maximizing the likelihood of observations and is not tied to the decision-making task. However, this approach can fail to find good policies when the model is (inevitably) misspecified; in particular, a maximum likelihood model may spend capacity modeling irrelevant information rather than signal important for the task at hand. We demonstrate this effect and show how POPCORN, which constrains maximum likelihood training of the POMDP model so that the value of the model's induced policy achieves satisfactory performance, can overcome these issues.

2 Related Work

RL in Healthcare. Healthcare applications of RL have proliferated in recent years, in diverse clinical areas such as schizophrenia (Shortreed et al., 2011),

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

[1] An implementation of POPCORN can be found online at `https://github.com/dtak/POPCORN-POMDP`, along with code to reproduce our experiments.

sepsis (Komorowski et al., 2018; Raghu et al., 2017), mechanical ventilation (Prasad et al., 2017), HIV (Ernst et al., 2006), and dialysis (Martín-Guerrero et al., 2009). However, most works use model-free approaches, ostensibly because learning accurate models from noisy biological data is challenging. All of these works further assume full-observability, which is often not accurate.

A few prior works in this applied space have explicitly modeled partial observability. POMDPs have been applied to heart disease management (Hauskrecht and Fraser, 2000), sepsis treatment in off-policy or simulated settings (Li et al., 2018; Oberst and Sontag, 2019; Peng et al., 2018), and HIV management (Parbhoo et al., 2017). All of these approaches take a two-stage approach to learning. In contrast, our approach is *decision-aware* throughout the optimization process.

Imperfect Models in RL Model-based RL is a long-standing area of research, and work as early as Abbeel et al. (2006) looked at learning misspecified models that are still useful for RL. More broadly, "end-to-end" optimization methods directly incorporate a downstream decision-making task during model training, and are growing in popularity across machine learning, from graphical models (Lacoste–Julien et al., 2011) to submodular optimization (Wilder et al., 2019). Within RL, recent decision-aware optimization efforts have explored partially-observed problems in both model-free (Karkus et al., 2017) and model-based settings (Igl et al., 2018).

These RL efforts differ from ours in two key respects. First, they exclusively focus on on-policy settings for simulated environments such as Atari. Second, they rely heavily on black-box neural networks for feature extraction, which are not generally sample-efficient or easily interpreted. In many cases (e.g. Karkus et al. (2017)), the model is treated as an abstraction and there is no way to set the importance of the model's ability to accurately generate trajectories. Perhaps closest in spirit to our approach is theoretical work on value-aware model learning in RL (Farahmand, 2018).

3 Background

POMDP Model. We consider a POMDP with K discrete latent states (e.g. physiological conditions of patients), A discrete actions (e.g. possible treatments), D-dimensional observations (e.g. clinical measurements), and deterministic rewards (e.g. how "good" or "bad" the treatments were). The entire generative model for states $s_t \in \{1, 2, \dots K\}$ and observations $o_t \in \mathbb{R}^D$ across timesteps $t \in \{0, 1, \dots T\}$ is given by:

$$\begin{aligned} p(s_0 = k) &\triangleq \tau_{0k}, \qquad (1)\\ p(s_{t+1} = k | s_t = j, a_t = a) &\triangleq \tau_{ajk}, \\ p(o_{t+1,d} | s_{t+1}{=}k, a_t{=}a) &\triangleq \mathcal{N}(\mu_{akd}, \sigma^2_{akd}). \end{aligned}$$

We define the model parameters as $\theta \equiv \{\tau, \mu, \sigma, R\}$. τ describes the transition probability of moving to the next (unobserved) state s_{t+1}, given current state s_t and action a_t. We model each observation o_t as Gaussian, with emission parameters μ and σ^2 denoting the mean and variance when in state s_t after taking action a_{t-1}. Although any (tractable) distribution is possible, we choose to use independent Gaussians across the D dimensions for simplicity. Completing the POMDP specification is the deterministic reward function $R(s, a)$, specifying the reward from taking action a in state s.

A dataset consists of N trajectories (e.g. decisions made about a patient's care, along with clinical observations). We index each trajectory by $n \in \{1, \dots, N\}$, with the length of trajectory n given by $T_n \leq T$.

Given a POMDP with parameters θ, we can compute the *belief* $b_t \in \Delta^K$, a vector in the simplex $\Delta^K \triangleq \{q \in \mathbb{R}^K | \sum_{k=1}^K q_k = 1, q_k \geq 0\}$. The belief defines the posterior over state s_t given all past actions and observations: $b_{tk} \triangleq p(s_t = k | o_{0:t}, a_{0:t-1})$, is a sufficient statistic for the entire history, and can be computed efficiently via forward filtering (Rabiner, 1989). We can *solve* the POMDP using a planning algorithm to learn a policy $\pi_\theta : \Delta^K \to \Delta^A$, mapping any belief to a distribution over actions (or a single action for deterministic policies). The goal is to find a policy with high value (the expected sum of discounted rewards): $V^\pi = \sum_{t=0}^T \gamma^t \mathbb{E}[r_t]$, with $\gamma \in (0, 1)$ the discount.

Learning Parameters: Input-Output HMM. The model in Eq. (1) is an input-output hidden Markov model (IO-HMM) (Bengio and Frasconi, 1995), where actions are inputs and observations are outputs. The model parameters $\{\tau, \mu, \sigma\}$ that maximize the marginal likelihood of observed trajectories can be efficiently computed using the EM algorithm for HMMs (Rabiner, 1989; Chrisman, 1992). For Bayesian approaches, efficient algorithms for sampling from the posterior over POMDP models also exist (Doshi-Velez, 2012). The deterministic reward function R is estimated separately by minimizing squared errors with the observed rewards (see Appendix B for additional details).

Solving for the Policy. The value function of a discrete-state POMDP can be modeled arbitrarily closely as the upper envelope of a finite set of linear functions of the belief (Sondik, 1978). However, exact value iteration is intractable even for very small POMDPs. In this work, we use point based value iteration (PBVI) (Pineau et al., 2003), an approximate algorithm that is significantly more efficient (see Shani et al. (2013) for a survey of related algorithms and extensions). Rather than perform Bellman backups over all valid beliefs $b \in \Delta^K$, PBVI only performs backups at a specific (small) set of beliefs. For the modest state spaces in our applications ($K << 100$), PBVI is an efficient solver. However, we require two key innovations beyond standard PBVI. First, we adapt ideas from Hoey and Poupart (2005) to handle continuous observations. Second, we relax the algorithm so that each step is differentiable. See Appendix A for details.

Off-Policy Value Estimation. Let π_θ be the (near-optimal) policy obtained from PBVI for the given model parameters $\theta = \{\tau, \mu, \sigma, R\}$. The fact that π_θ is optimal for a specific model θ does not mean it is optimal in practice (e.g. in the clinic), because our generative model is almost certainly misspecified. If we have access to an environment simulator, we can evaluate π_θ via standard Monte Carlo rollouts. However, in the *batch* setting, we lack the ability to interact with the environment and must turn to *off policy evaluation* (OPE) to *estimate* a policy's value.

Let π_{beh} denote the behavior policy under which the observed data were collected (e.g. clinician treatment tendencies).[2] Let $\mathcal{D}$ denote a set of N trajectories collected under π_{beh}. In this work, the specific OPE technique we use is consistent weighted per-decision importance sampling (CWPDIS, (Thomas, 2015)) which estimates the value of a policy π_θ as:

$$V^{\text{CWPDIS}}(\pi_\theta) \triangleq \sum_{t=1}^{T} \gamma^t \frac{\sum_{n\in\mathcal{D}} r_{nt}\rho_{nt}(\pi_\theta)}{\sum_{n\in\mathcal{D}} \rho_{nt}(\pi_\theta)}, \qquad (2)$$

$$\rho_{nt}(\pi_\theta) \triangleq \prod_{s=0}^{t} \frac{\pi_\theta(a_{ns}|o_{n,0:s}, a_{n,0:s-1})}{\pi_{\text{beh}}(a_{ns}|o_{n,0:s}, a_{n,0:s-1})}.$$

In general, importance sampling (IS) estimators such as CWPDIS have lower bias than other approaches but suffer from higher variance. Another class of OPE methods learn a separate model to simulate trajectories in order to estimate policy values (e.g. Chow et al. (2015)), but may suffer from unacceptably high bias in real-world, noisy settings.

4 Prediction-Constrained POMDPs

We now introduce POPCORN, our proposed *prediction-constrained* optimization framework for learning POMDPs. We seek to learn parameters θ that will both assign high likelihood to the observed data while also yielding a policy π_θ with high (estimated) value. As noted in Sec. 2, previous approaches for learning POMDPs generally fall into two categories. Two-stage methods (e.g. Chrisman (1992)) that first learn a model and then solve it often fail to find good policies under severe model misspecification. End-to-end methods (e.g. Karkus et al. (2017)) that focus only on the "discriminative" task of policy learning typically fail to produce accurate generative models of the environment. They also lack the ability to handle missing observations, which is especially problematic in medical contexts where missing data is pervasive.

Our approach offers a *balance* between these purely maximum likelihood-driven (generative) and purely reward-driven (discriminative) extremes. We retain the strengths of the generative approach—the ability to plan under missing observations, simulate accurate dynamics, and inspect model parameters to inspire scientific hypotheses—while benefiting from model parameters that are directly informed by the decision task as in end-to-end frameworks.

4.1 POPCORN Objective

Our proposed framework seeks a POMDP θ that maximizes the log marginal likelihood $\mathcal{L}_{\text{gen}}$ of the observed data $\mathcal{D}$, while enforcing that the solved policy's (estimated) value $V(\pi_\theta)$ is high enough to be useful. Formally, we seek a θ that maximizes the constrained optimization problem:

$$\max_\theta \mathcal{L}_{\text{gen}}(\theta), \qquad \text{subject to: } V(\pi_\theta) \geq \epsilon, \qquad (3)$$

with the functions $\mathcal{L}_{\text{gen}}$ and V defined below. The tolerance ϵ defines a minimum acceptable policy value (e.g. as determined by a domain expert).

Setting practical optimization considerations aside, we would prefer the constrained formulation of Eq. (3) as it best expresses our model-fitting goals: as good a generative model as possible, but we will not accept poor decision-making. This objective is similar to the *prediction-constrained* objective used by Hughes et al. (2018) for optimizing supervised topic models; here we apply similar ideas to batch, off-policy RL settings.

In practice, solving constrained problems is challenging, so we transform to an equivalent unconstrained objective using a Lagrange multiplier $\lambda > 0$:

$$\max_\theta \quad \mathcal{L}_{\text{gen}}(\theta) + \lambda V(\pi_\theta). \qquad (4)$$

Setting $\lambda = 0$ recovers classic two-stage training, while the limit $\lambda \to \infty$ approximates end-to-end approaches. In our experiments, we compare against both of these baseline approaches, referring to the $\lambda = 0$ case as "2-stage", and the $\lambda \to \infty$ case as "Value-only" (for this case, in practice we optimize $V(\pi_\theta)$ and ignore $\mathcal{L}_{\text{gen}}$).

Computing the Generative Term. We define $\mathcal{L}_{\text{gen}}(\theta)$ as the log marginal likelihood of observations, given the actions in $\mathcal{D}$ and parameters θ:

$$\mathcal{L}_{\text{gen}}(\theta) = \textstyle\sum_{n\in\mathcal{D}} \log p(o_{n,0:T_n}|a_{n,0:T_n-1}, \theta). \qquad (5)$$

This IO-HMM likelihood marginalizes over uncertainty about the hidden states, can be computed efficiently via dynamic programming (Rabiner, 1989), and is also amenable to automatic differentiation w.r.t. θ.

[2] In our experiments with simulated environments, we assume π_{beh} is given. In the real data setting, we estimate the behavior policy via the k-nearest neighbors approach of Raghu et al. (2018). See Appendix E.4 for details.

Computing the Value Term. Computation of $V(\pi_\theta)$ entails two distinct parts: solving for the policy π_θ given θ, and then estimating the value of this policy using OPE and $\mathcal{D}$. We require both to be differentiable to permit gradient-based optimization. To solve for the policy, we apply a differentiable relaxation of PBVI (see Appendix A.4 for full details). Although standard PBVI returns a deterministic policy, we relax this as well and learn stochastic policies as they are generally easier to evaluate with OPE. We emphasize that our framework is general and other solvers are possible as long as they can be made differentiable. To compute the estimated policy value, we use the CWPDIS estimator in Eq. (2). As it is a *differentiable* function of θ, our unconstrained objective in Eq. (4) can be optimized via first-order gradient ascent.

4.2 Optimizing the Objective

We optimize using gradients computed from the full dataset (we do not use subsample to avoid extra variance). We optimize with Rprop (Igel and Hüsken, 2003) with default settings. Our objective is challenging due to non-convexity, as even the generative term alone admits many local optima. To improve solution quality in all experiments and for all methods, prior to final evaluation we take the best of 25 random restarts as measured by training objective value.

Stabilizing the Off-Policy Estimate. Although our OPE estimate (using CWPDIS) was reliable in simulated environments, on our real dataset it had unusably high variance, as is common with IS estimators. We address this in two ways.

First, we add an extra term in the objective encouraging larger effective sample size (ESS) and hence lower variance, following Metelli et al. (2018). Our final objective includes an ESS penalty with weight $\lambda_{\text{ESS}} > 0$:

$$\max_{\theta} \quad \mathcal{L}_{\text{gen}}(\theta) + \lambda \cdot \left[V(\pi_\theta). - \frac{\lambda_{\text{ESS}}}{\sqrt{\text{ESS}(\theta)}} \right]. \qquad (6)$$

As the CWPDIS estimator in Eq. 2 is the weighted sum of a sequence of T IS estimators (the average discounted reward at each t), we sum all these stepwise ESS_t values to yield the final $\text{ESS}(\theta)$ term. ESS_t is approximated given IS weights $\{\rho_{nt}\}_{n=1}^N$ as $\frac{(\sum_n \rho_{nt})^2}{\sum_n \rho_{nt}^2}$ (Kong, 1992).

Second, we restrict the support of π_θ and then renormalize to only allow actions where there was at least δ probability under π_{beh}. This forces strong overlap between the support of π_{beh} and π_θ and often substantially reduces the variance of the final OPE estimate. This also provides a soft notion of "safety", as now rare or unknown actions are prohibited.

Hyperparameters. The key hyperparameter for POPCORN is the scalar tradeoff $\lambda > 0$. We try a range of 5 λ's per environment spaced evenly on a log scale. We also rescale $\mathcal{L}_{\text{gen}}(\theta)$ by the total number of observed scalars ($D(\sum_n T_n)$ if there is no missing data), so that the magnitude of λ has roughly consistent impact across datasets. We also try 5 ESS penalty weights λ_{ESS} evenly spaced on a log scale, but this term was only necessary for the real data experiments.

5 Simulated Environments

We first evaluate POPCORN on three simulated environments to validate its utility across a range of possible model misspecification scenarios. We later evaluate on a more difficult medical simulator. For all experiments in this section, everything is conducted in the batch, off-policy setting. The simulator is only used to produce the initial data set and to evaluate the final policy after training concludes. We separate each experiment into a description of procedure and highlights of the results.

Recall that our goal is to learn simple—and therefore interpretable—models that perform robustly in misspecified settings. As such, we compare against an approach that does not attempt to model the dynamics ("value term only"), an approach that first learns the model and then plans ("2-stage"), and a known optimal solution (when available). In all cases, we are interested in how these methods trade off between explaining the data well (log marginal likelihood of data) and making good decisions (policy value).

5.1 Synthetic Domains with Misspecification

We demonstrate how POPCORN overcomes various kinds of model misspecification in the context of the classic POMDP tiger problem (Kaelbling et al., 1998). The tiger problem consists of a room with 2 doors: one door is safe, and the other door has a tiger behind it. The agent has 3 possible actions: either open one of the doors, thereby ending the episode, or listen for noisy evidence of which door is safe to open. Revealing a tiger gives -5 reward, the safe door yields $+1$ reward, and listening incurs -0.1 reward. The goal is to maximize rewards over many repeated trials, with the "safe" door's location randomly chosen each time.

We set $\gamma = 0.9$ to encourage the agent to act quickly. We collect training data from a random policy that first listens for 5 time steps, and then randomly either opens a door or listens again. We train in the *batch setting* given a single collection of 1000 trajectories of length 5-15. After optimization, we evaluate each policy via an additional 1000 Monte Carlo rollouts.

Tiger with Irrelevant Noise: Finding dimensions that signal reward. In this setting, whenever the agent listens for the tiger, it receives an observation o_t with $D = 2$ dimensions. The first dimension provides a noisy signal as to the location of the safe door. We set this "signal" dimension $o_{t1} \sim \mathcal{N}(i_{\text{safe}}, 0.3^2)$, where the

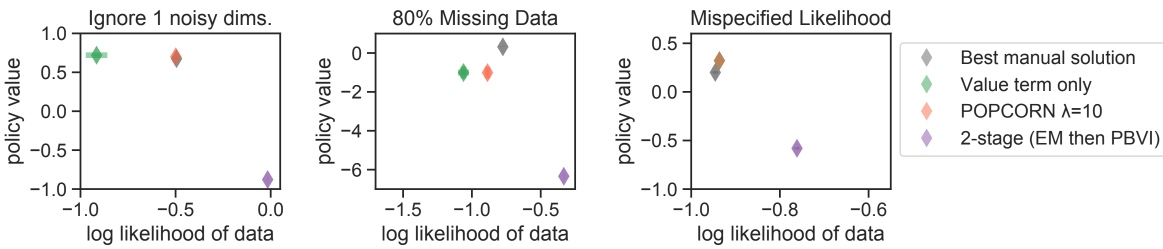

Figure 1: Solutions from all competitor methods in the 2D fitness landscape (policy value on y-axis; log marginal likelihood on x-axis). An ideal method would score in the top right corner of each plot. *Left*: Results from Tiger with Irrelevant Noise Dimensions. *Middle*: Results from Tiger with Missing Data. *Right*: Results from Tiger with Wrong Likelihood. POPCORN is robust to all three types of model misspecification tested, and consistently learns better policies than 2-stage and better models than value-only.

mean is the safe door's index $i_{\text{safe}} \in \{0, 1\}$. The second dimension is irrelevant to the safe door's location, and we set $o_{t2} \sim \mathcal{N}(j, 0.1^2)$, with $j \sim \text{Unif}(\{0, 1\})$ in each trial. Thus, $K = 4$ total states would be needed to explain perfectly both the relevant and irrelevant signals for all possible values of (i_{safe}, j).

We measure performance allowing only $K = 2$ states to assess how each method spends its limited capacity across the generative and reward-seeking goals. We expect the 2-stage baseline will identify the irrelevant states indexed by j, as they have lower standard deviation (0.1 vs. 0.3 for the signal dimension) and thus are more important to maximize likelihood. In contrast, we expect POPCORN will focus on the relevant signal dimension and recover high-value policies.

Tiger with Missing Data: Finding relevant dimensions when some data is missing. This next scenario extends the previous setting in which the listen action produces $D = 2$ observations, where the first signals the safe door's location and the second is irrelevant. However, now the dimension with the relevant signal is often missing. Specifically, $o_{t1} \sim \mathcal{N}(i_{\text{safe}}, 0.3^2)$ and $o_{t2} \sim \mathcal{N}(j, 0.3^2)$, but we select 80% of signal observations o_{t1} to be *missing* uniformly at random. This (coarsely) simulates clinical settings where some measurements may be infrequent but important (e.g. relevant lab tests), while others are common but not directly useful (e.g. routine vitals).

The expected outcome with $K = 2$ states is that a 2-stage approach driven by maximizing likelihood would prefer to model the always-present irrelevant dimension. In contrast, POPCORN should learn to favor the signal dimension even though it is rarely available and contributes less overall to the likelihood. This ability to gracefully handle missing dimensions is a natural property of generative models and would not be easily done with a model-free approach.

Tiger with Wrong Likelihood: Overcoming a misspecified density model. Finally, we consider a situation in which our generative model's *density family* cannot match the true observation distribution. This time, the listen action produces a $D = 1$ dimensional observation o_t. The true distribution of this observation signal is a *truncation* of a mixture of two Gaussians, denoted $\text{GMM}(o) = 0.5\mathcal{N}(o|0, 0.1^2) + 0.5\mathcal{N}(o|1, 1.0^2)$. If the first door is safe, listening results in strictly negative observations: $p(o) \propto \delta(o < 0)\text{GMM}(o)$. If the second door is safe, listening results in strictly positive observations: $p(o) \propto \delta(o > 0)\text{GMM}(o)$.

While the the true observation densities are *not* Gaussian, we will fit POMDP models with Gaussian likelihoods and $K = 2$ states. We expect POPCORN to still deliver high-value policies, even though the likelihood will be suboptimal. See Appendix C for more details on the overall setup of all three tiger environments as well as additional results.

5.2 Conclusions from Synthetic Domains

Across all variants of the Tiger problem, we observe many common conclusions from Fig. 1. Together, these results demonstrate how POPCORN is robust to many different kinds of model misspecification.

POPCORN learns consistently better policies than 2-stage. Across all 3 panels of Fig. 1, POPCORN (red) delivers higher value $V(\pi_\theta)$ (y-axis) than the 2-stage baseline (purple).

Value-only learns poor generative models. In 2 of 3 panels, the value-only baseline (green) has noticeably worse likelihood $\mathcal{L}_{\text{gen}}(\theta)$ (x-axis) than POPCORN. The far right panels shows indistinguishable performance. Notably, optimizing this objective is significantly less stable than the full POPCORN objective. This aligns with findings from Levine and Koltun (2013), who also observed that policy learning via direct optimization of IS estimates of policy value is challenging.

POPCORN solutions are consistent with manually-designed solutions. In all 3 panels, POPCORN (red) is the closest method to the ideal manually-designed solution (gray).

5.3 Sepsis Simulator: Medically-motivated environment with known ground truth.

We now move from simple toy problems—each designed to demonstrate a particular robustness of our method—to a more challenging simulated domain. In real-world medical decision-making tasks, it is impossible to evaluate the value of a learned policy using data collected under that policy's decisions. However, in a simulated setting, we can evaluate any given policy to assess its true value. We emphasize θ is still learned in the batch setting, as only after optimization do we use the simulator to allow for accurate evaluation of policy values.

We use the simulator from Oberst and Sontag (2019), which is a coarse physiological model for sepsis with $D = 5$ discrete observations: 4 ordinal-valued vitals (e.g. "low"/"normal"/"high"), and a binary indicator for diabetes. The true simulator is governed by an underlying known Markov decision process (MDP), which has 1440 possible discrete states. There are 8 actions (3 different binary actions), and trajectories are at most 20 timesteps. Rewards are sparse, with 0 reward at intermediate time steps and -1 or $+1$ at termination.

To make this simulator similar to our other environments with continuous-valued observations, we add independent Gaussian noise with standard deviation 0.3 to each observation. This measurement error also makes the environment partially observable so that modeling it as a POMDP is reasonable. Although Oberst and Sontag (2019) used structural causal models to simulate counterfactual trajectories and explicitly address causal questions, our POMDP construction implicitly assumes no hidden confounding. Our work skirts causality, as we view POMDPs solely as a convenient way to summarize trajectory histories. Our use of this simulator hence differs substantially from its original use, where it was used to create strong (known) hidden confounding in order to illustrate failure modes of OPE[3].

The true discrete-state MDP is easily solved via exact value iteration. We generate 2500 trajectories under an ϵ-greedy behavior policy, with $\epsilon = 0.14$ so each non-optimal action has a .02 probability of being taken. Given observed trajectories, we learn POMDPs assuming $K = 5$ (we obtained similar qualitative results for other K), and evaluate policies via an additional 2500 Monte Carlo rollouts. See Appendix D for full details.

Results and Conclusions. Figure 2 shows POPCORN again learns higher-value policies than 2-stage. Additionally, while the value-term-only baseline learns a policy on par with POPCORN, its likelihood is substantially lower. While all policies are far from optimal, this unsurprising given the small state space, modest observation noise, and high ϵ for the behavior policy.

[3]We refer interested readers to recent related work in Tennenholtz et al. (2020) that specifically proposes a new technique for performing OPE with POMDPs where there is hidden confounding. This differs from Oberst and Sontag (2019) who focus solely on a new approach to generating counterfactual trajectories under a proposed policy.

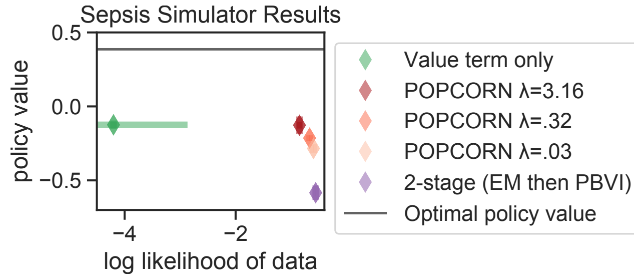

Figure 2: Quantitative results from the sepsis simulator. An ideal method would score in the top right corner. No methods recover a fully optimal policy (grey line), but POPCORN consistently learns better policies than 2-stage and better models than value-only. A policy which takes actions uniformly at random has a value of -0.72, so the 2-stage policy barely outperforms this.

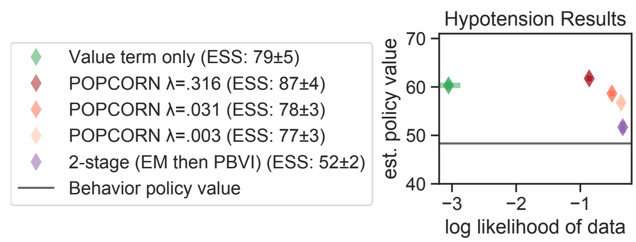

Figure 3: Quantitative results from the hypotension data, showing the trade-offs between policy value and likelihood (ESS is in legend). POPCORN again learns much better models than value-only and better policies than 2-stage.

6 Real Data Application: Hypotension

To showcase the utility of our method on a real-world medical decision making task, we apply POPCORN to the challenging problem of managing acutely hypotensive patients in the ICU. Although hypotension is associated with high morbidity and mortality (Jones et al., 2006), management of these patients is difficult and treatment strategies are not standardized, in large part because there are many underlying potential causes of hypotension. Previously, Girkar et al. (2018) attempted to predict the efficacy of fluid therapy for hypotensive patients with only mixed success. We apply POPCORN to this problem and first study the same trade-offs between generative and reward-seeking performance as in Sec. 5. We further perform an in-depth evaluation of the learned policy and our confidence in it (via effective sample sizes and qualitative checks).

Cohort. We use 10,142 ICU stays from MIMIC-III (Johnson et al., 2016), filtering to adult patients with at least 3 abnormally low mean arterial pressure (MAP) values in the first 72 hours of ICU admission. Our observations consist of 9 vitals and laboratory measurements: MAP, heart rate, urine output, lactate, Glasgow coma score, serum creatinine, FiO_2, total bilirubin, and platelets count. We discretized time into 1-hour windows, and setup the RL task to begin 1 hour after ICU admission to ensure a sufficient amount of data exists before starting a policy. Trajectories end either at ICU discharge or at 72 hours into the ICU admission, so there are at most 71 actions taken. This formulation was made in consultation with a critical care physician, who advised most acute cases of hypotension would present early during an ICU admission. We expressly do *not* impute missing observations: only observed measurements contribute to the overall likelihood.

Setup. Our action space consists of the two main treatments for acute hypotension: fluid bolus therapy and vasopressors, both of which are designed to quickly raise blood pressure and increase perfusion to the organs. We discretize fluids into 4 actions (none, low, medium, high), and discretize vasopressors into 5 actions (none, low, medium, high, very high) for a total of 20 discrete actions. To assign rewards to individual time steps, we use a piecewise-linear function created in consultation with a critical care physician. A MAP of 65mmHg is a common target (Asfar et al., 2014), so if an action is taken and the next MAP is 65 or higher, the next reward is $+1$, the highest possible value. Otherwise, rewards decrease as MAP values drop, with MAP ≤ 30 delivering a reward of 0, the smallest possible value. Further details on the action space discretization, a plot of the reward function, and other preprocesssing can be found in Appendix E.

We split the dataset into 5 distinct test sets for cross-validation, and throughout present results on the test sets, with standard errors across folds where appropriate. We set $\lambda_{\text{ESS}} = 4$ and set $\delta = 0.03$, which prohibits all actions assigned less than 3% probability by our estimated behavior policy. Lastly, we study several possible values for the Lagrange multiplier, $\lambda \in \{10^{-2.5}, 10^{-1.5}, 10^{-0.5}\}$.

6.1 Conclusions from ICU Application

POPCORN achieves the best balance of high-performing policies and high likelihood models. As in earlier results, Figure 3 shows how POPCORN balances generative and decision-making performance well, with darker red indicating higher λ's and thus improved policy values. The policy values for the 2-stage baseline and the likelihood scores for the value-only baseline both substantially underperform POPCORN.

POPCORN has reasonably accurate forecasts. To demonstrate the ability of models to predict future observations, Figure 4 shows results from a forecasting experiment. Each method is given the first 12 hours of a trajectory, and then must predict future observations up to 12 hours in the future. Importantly, only *measured* observations are used to calculate the mean absolute error between model predictions and true values. Unsurprisingly 2-stage generally performs the best, although POPCORN for small values of λ often performs similarly. On the other hand, the value-only baseline fares significantly worse. For some observations (MAP and urine output; see left-most column of Figure 4), it makes nonsensical predictions far outside the range of observed data, with errors several orders of magnitudes worse than POPCORN and 2-stage.

POPCORN enables inspection if learned models are clinically sensible. We visualize the learned emission distributions for MAP across the $K = 5$ states and 20 actions for each method in Figure 5. Note that densities may appear non-Gaussian, as they are back-transformed to the original scale of the data but were modeled on a log-scale. POPCORN's distributions are more spread out and better differentiate between states compared to the 2-stage baseline, which learns very similar states with high overlap. As a result, the 2-stage policy will end up recommending similar actions for most patients. Value-only learns states that are even more diverse, allowing it to learn an effective policy but at the expense of not modeling the observed data well. See Appendix E.5 for similar results for lactate, urine output, and heart rate. Although these results are exploratory, these simple visualizations of what the models have learned are only possible due to the white-box nature of our HMM-based approach, compared with e.g. deep reinforcement learning methods.

Figure 6 visualizes the action probabilities for the behavior policy, a value-only policy, a POPCORN policy, and a 2-stage policy. In general, the POPCORN policy most closely aligns with the behavior, although it is also quite similar to value-only. On the other hand, the 2-stage policy seems in general more conservative and tends to have lower probabilities on more aggressive actions. In future work we plan to work with clinical partners to explore individual patient trajectories and understand how and why these treatment policies differ.

POPCORN learns models that transfer to other tasks. Figure 7 shows results testing how well models transfer to solving a new task. We use a new reward function that penalizes high lactate values (see Appendix E.3 for a plot). For each method, we freeze τ, μ, σ from the previous optimization, but learn a new R. Then we solve these new models to learn new policies, and estimate their values. We find that the POPCORN and two-stage models transfer reasonably well, whereas value-only is substantially worse especially given its high original estimated value in Figure 3.

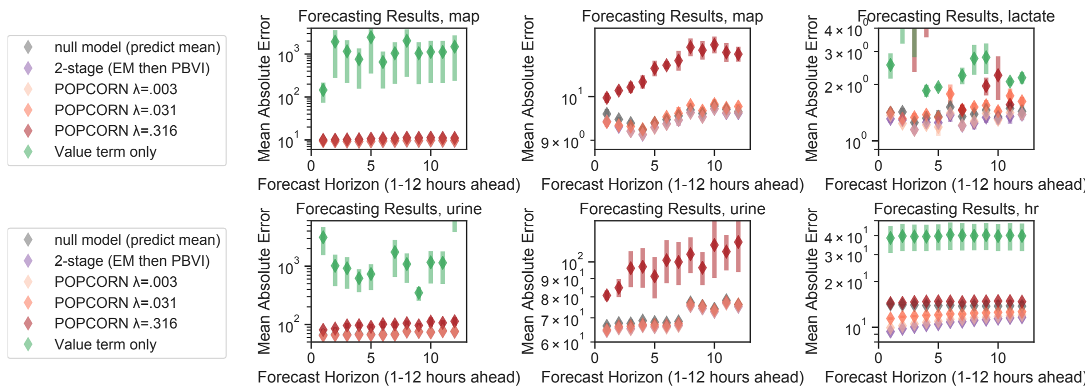

Figure 4: Forecasting results. *Top to bottom, left to right*: MAP (scale zoomed out); MAP (value-only out of pane); lactate; urine output (scale zoomed out); urine output (value-only out of pane); heart rate. 2-stage performs the best throughout, but for smaller values of λ POPCORN is often not much worse. Value-only constantly makes wildly inaccurate predictions, as its forecast errors are often several orders of magnitude worse (see MAP and urine results in first column).

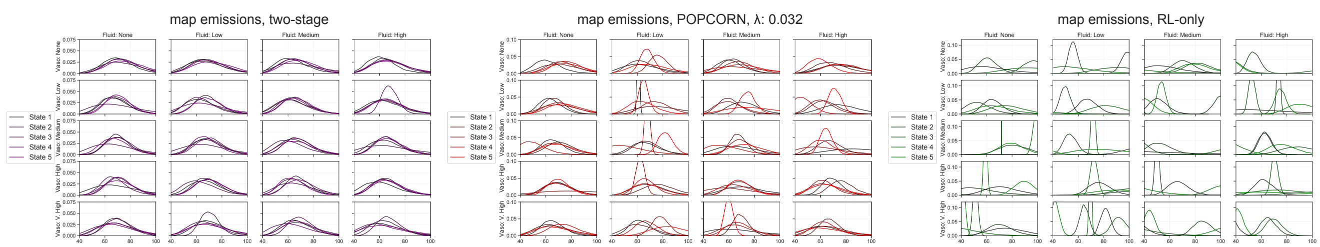

Figure 5: Visualization of learned MAP distributions. *Left:* 2-stage. *Middle:* POPCORN, $\lambda = 0.032$. *Right:* Value-only. Each subplot visualizes all 100 learned distributions of MAP values for a given method, across 20 actions and $K = 5$ states. Each pane in a subplot corresponds to a different action, and shows distributions across the 5 states. Vasopressors vary across rows, and fluids vary across columns. 2-stage learns states that are mostly homogeneous, value-only learns states that are differentiated and often far apart, while POPCORN is somewhere in between.

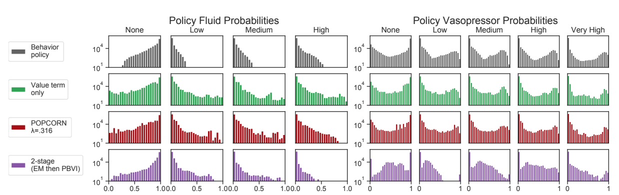

Figure 6: Action probabilities for the behavior policy, a value-only policy, a POPCORN policy with $\lambda = 0.316$, and a 2-stage policy. Actions are split from the full 20-dimensional space by type. *Left:* Action probabilities for the 4 doses of IV fluids, and *Right:* for the 5 doses of vasopressors.

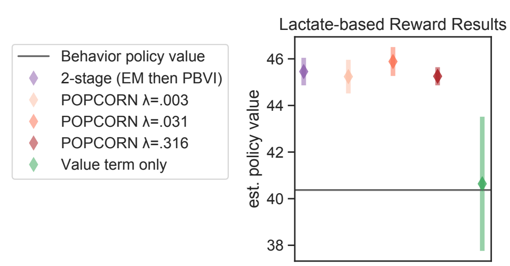

Figure 7: Results from reusing previously learned models from each approach, and solving them to learn new policies given a different reward function based on lactate values. 2-stage and POPCORN learn models that are able to transfer reasonably well to the new task, while value-only does not.

7 Discussion

We proposed POPCORN, an optimization objective for off-policy batch RL with partial observability. POPCORN balances the trade-off between learning a model with high likelihood and a model well-suited for planning, even in batch off-policy settings. Synthetic experiments demonstrate POPCORN achieves good policies and decent models even in the face of misspecification (in the number of states, the choice of the likelihood, or the availability of data). Performance on a clinical decision-making task suggests we may be able to learn a policy on par or even slightly better than the observed clinician behavior policy. Future directions include scaling to environments with more complex state structures or long-term temporal dependencies, investigating semi-supervised settings where not all sequences have rewards, better leveraging that the behavior policy is not terribly sub-optimal, and learning Pareto-optimal policies that balance multiple competing goals. We hope methods such as ours ultimately become integrated into clinical decision support tools to assist physicians in improving the treatment of critically ill patients.

Acknowledgements

FDV and JF acknowledge support from NSF Project 1750358. JF additionally acknowledges Oracle Labs, a Harvard CRCS fellowship, and a Harvard Embedded EthiCS fellowship. MCH acknowledges support from NSF Project HDR-1934553. The authors also thank David Sontag, Omer Gottesman, Leo Anthony Celi, Ryan Kindle, and the anonymous reviewers for thoughtful and constructive feedback.

References

M. I. M. García, P. G. González, M. G. Romero, A. G. Cano, C. Oscier, A. Rhodes, R. M. Grounds, and M. Cecconi. Effects of fluid administration on arterial load in septic shock patients. *Intensive care medicine*, 41(7):1247–1255, 2015.

O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25, 2019.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

S. M. Shortreed, E. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy. Informing sequential clinical decision-making through reinforcement learning: An empirical study. *Machine learning*, 84(1-2): 109–136, 2011.

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, 2018.

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. In *Machine Learning for Healthcare Conference (MLHC)*, pages 147–163, 2017.

N. Prasad, L. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In *Conference on Uncertainty in Artifical Intelligence (UAI)*, 2017.

D. Ernst, G. Stan, J. Goncalves, and L. Wehenkel. Clinical data based optimal STI strategies for HIV: A reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672. IEEE, 2006.

J. D. Martín-Guerrero, F. Gomez, E. Soria-Olivas, J. Schmidhuber, M. Climente-Martí, and N. V. Jiménez-Torres. A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert Systems with Applications*, 36(6):9737–9742, 2009. ISSN 0957-4174.

M. Hauskrecht and H. Fraser. Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine*, 18(3):221–244, 2000.

L. Li, M. Komorowski, and A. A. Faisal. The actor search tree critic (ASTC) for off-policy POMDP learning in medical decision making. *arXiv preprint arXiv:1805.11548*, 2018.

M. Oberst and D. Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *International Conference on Machine Learning (ICML)*, 2019.

X. Peng, Y. Ding, D. Wihl, O. Gottesman, M. Komorowski, L. H. Lehman, A. Ross, A. Faisal, and F. Doshi-Velez. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, page 887. American Medical Informatics Association, 2018.

S. Parbhoo, J. Bogojeska, M. Zazzi, V. Roth, and F. Doshi-Velez. Combining kernel and model based learning for HIV therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

P. Abbeel, M. Quigley, and A. Ng. Using inaccurate models in reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2006.

S. Lacoste–Julien, F. Huszár, and Z. Ghahramani. Approximate inference for the loss-calibrated Bayesian. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

B. Wilder, B. Dilkina, and M. Tambe. Melding the data-decisions pipeline: decision-focused learning for combinatorial optimization. In *AAAI Conference on Artificial Intelligence*, 2019.

P. Karkus, D. Hsu, and W. S. Lee. QMDP-Net: Deep learning for planning under partial observability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson. Deep variational reinforcement learning for POMDPs. In *International Conference on Machine Learning (ICML)*, 2018.

A. Farahmand. Iterative value-aware model learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9072–9083, 2018.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

Y. Bengio and P. Frasconi. An input output HMM architecture. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1995.

L. Chrisman. Reinforcement learning with perceptual aliasing: the perceptual distinctions approach. In *AAAI Conference on Artificial Intelligence*, 1992.

F. Doshi-Velez. *Bayesian nonparametric approaches for reinforcement learning in partially observable domains*. PhD thesis, Massachusetts Institute of Technology, 2012.

E. J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: discounted costs. *Operations Research*, 26(2), 1978.

J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: an anytime algorithm for POMDPs. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2003.

G. Shani, J. Pineau, and R. Kaplow. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51, 2013.

J. Hoey and P. Poupart. Solving POMDPs with continuous or large discrete observation spaces. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

A. Raghu, O. Gottesman, Y. Liu, M. Komorowski, A. A. Faisal, F. Doshi-Velez, and E. Brunskill. Behaviour policy estimation in off-policy policy evaluation: calibration matters. *arXiv preprint arXiv:1807.01066*, 2018.

P. S. Thomas. *Safe reinforcement learning*. PhD thesis, University of Massachusetts, Amherst, 2015.

Y. Chow, M. Petrik, and M. Ghavamzadeh. Robust policy optimization with baseline guarantees. *arXiv preprint arXiv:1506.04514*, 2015.

M. C. Hughes, G. Hope, L. Weiner, T. H. Mccoy, R. H. Perlis, E. Sudderth, and F. Doshi-Velez. Semi-supervised prediction-constrained topic models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

C. Igel and M. Hüsken. Empirical evaluation of the improved rprop learning algorithms. *Neurocomputing*, 50:105–123, 2003.

A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

A. Kong. A note on importance sampling using standardized weights. Technical Report 348, University of Chicago Department of Statistics, 1992.

S. Levine and V. Koltun. Guided policy search. In *International Conference on Machine Learning (ICML)*, 2013.

G. Tennenholtz, S. Mannor, and U. Shalit. Off-policy evaluation in partially observable environments. In *AAAI Conference on Artificial Intelligence*, 2020.

A. E. Jones, V. Yiannibas, C. Johnson, and J. A. Kline. Emergency department hypotension predicts sudden unexpected in-hospital mortality: A prospective cohort study. *CHEST*, 130(4):941–946, 2006.

U. M. Girkar, R. Uchimido, L. H. Lehman, P. Szolovits, L. A. Celi, and W. Weng. Predicting blood pressure response to fluid bolus therapy using attention-based neural networks for clinical interpretability. *arXiv preprint arXiv:1812.00699*, 2018.

A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, 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.

P. Asfar, F. Meziani, J. Hamel, F. Grelon, B. Megarbane, N. Anguel, J. Mira, P. Dequin, S. Gergaud, and N. Weiss. High versus low blood-pressure target in patients with septic shock. *New England Journal of Medicine*, 370(17):1583–1593, 2014.