POPCORN: Partially Observed Prediction Constrained Reinforcement Learning

AUTHORS: JOSEPH FUTOMA, MICHAEL C. HUGHES, FINALE DOSHI-VELEZ

---

presenter: Zhongwen Zhang

CS885 – University of Waterloo – July, 2020

Overview

- Problem: ***decision-making*** for managing patients in ICU (Intensive Care Unit) with acute hypotension

- Challenges:
  - Medical environment is partially observable → POMDP
  - Model misspecification → POPCORN
  - Data limited → OPE (Off Policy Evaluation)
  - Data missing → Generative model

Solutions:

- Importance: more effective treatment is badly needed

Related work

- Model-free RL methods assuming full-observability [Komorowski et al., 2018] [Raghu et al., 2017] [Prasad et al., 2017] [Ernstet al., 2006] [Martín-Guerrero et al., 2009].

- POMDP RL methods (two-stage fashion) [Hauskrecht and Fraser, 2000] [Li et al., 2018] [Oberst and Sontag, 2019]

- Decision-aware optimization:
  - Mode-free [Karkus et al., 2017]
  - Model-based [Igl et al., 2018]

1. On-policy setting
2. Features extracted from network

High-level Idea

- Find a balance between purely maximum likelihood estimation (generative model) and purely reward-driven (discriminative model) extreme.

Prediction-Constrained POMDPs

- Objective:

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

- Equivalently transformed objective:

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

- Optimization method: gradient descent

Log Marginal Likelihood $\mathcal{L}_{gen}$

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

- Computation: EM algorithm for HMM [Rabiner, 1989]
- Parameter set: $\theta \equiv \{\tau, \mu, \sigma, R\}$

Estimated separately

$$p(s_0 = k) \triangleq \tau_{0k},$$

$$p(s_{t+1} = k | s_t = j, a_t = a) \triangleq \tau_{ajk},$$

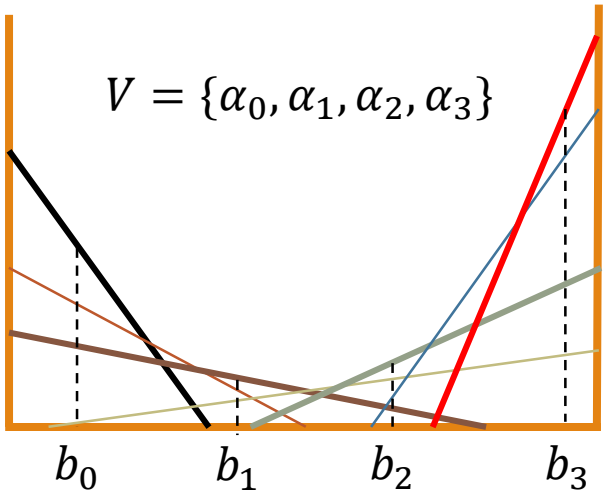
$$p(o_{t+1,d} | s_{t+1} = k, a_t = a) \triangleq \mathcal{N}(\mu_{akd}, \sigma^2_{akd})$$

Computing the value term $V(\pi_\theta)$

- Step1: Computing π_θ by PBVI (Point-Based Value Iteration)
- Step2: Computing $V(\pi_\theta)$ by OPE

Computing the value term $V(\pi_\theta)$

- Step1: Computing π_θ by PBVI (Point-Based Value Iteration)
  [Joelle Pineau, et.al., 2003]
  - Exact value iteration costs **exponential** time complexity
  - Approximation by only computing the value for a set of belief points
    polynomial time complexity



Computing the value term $V(\pi_\theta)$

- Step1: Computing π_θ by PBVI (Point-Based Value Iteration)
- Step2: Computing $V(\pi_\theta)$ by OPE
  - π_θ vs. $\pi_{behavior}$
  - Importance sampling: $\mathbf{E}_q\left[\frac{p(X)}{q(X)}f(X)\right] = \mathbf{E}_p[f(X)]$ → $IS(D) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{L}\frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}\right)\left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$

    under some mild assumption
    - Lower bias
    - Sample efficient

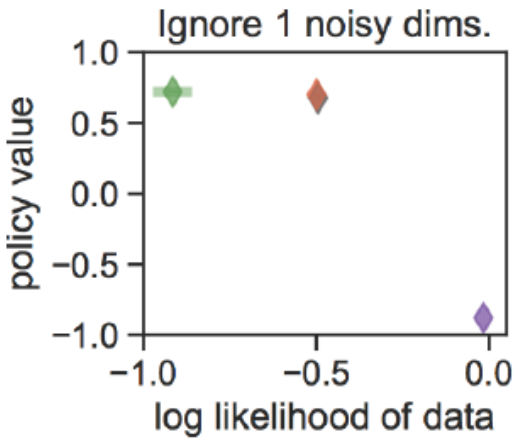
Empirical evaluation

- Simulated environments
  - Synthetic domain
  - Sepsis simulator
- Real data application: hypotension

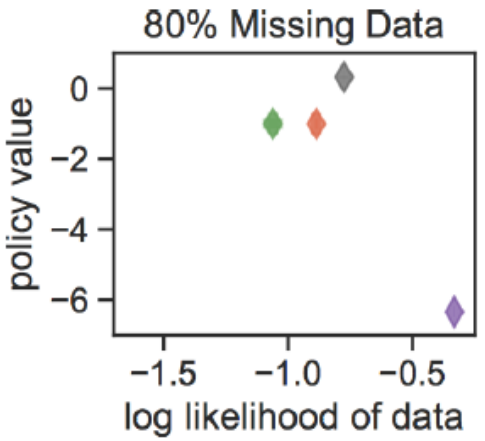
Synthetic domain

problem setting:

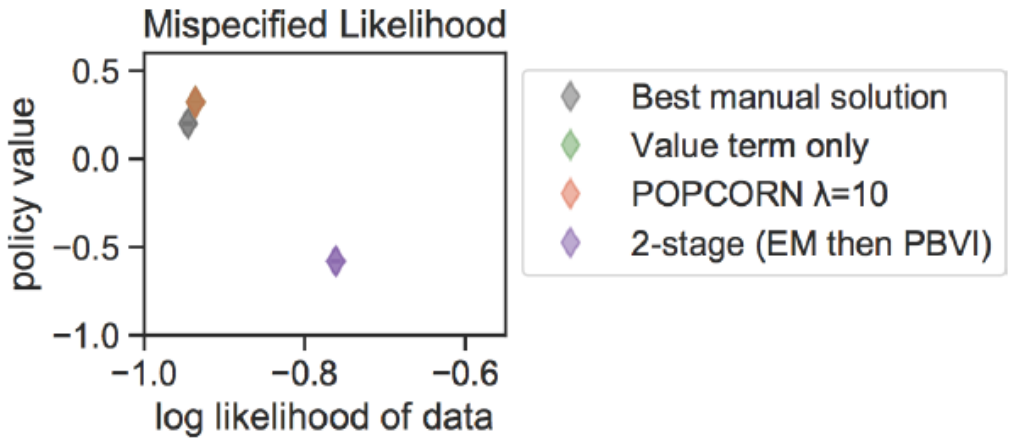
Synthetic domain



finding relevant signal dimension



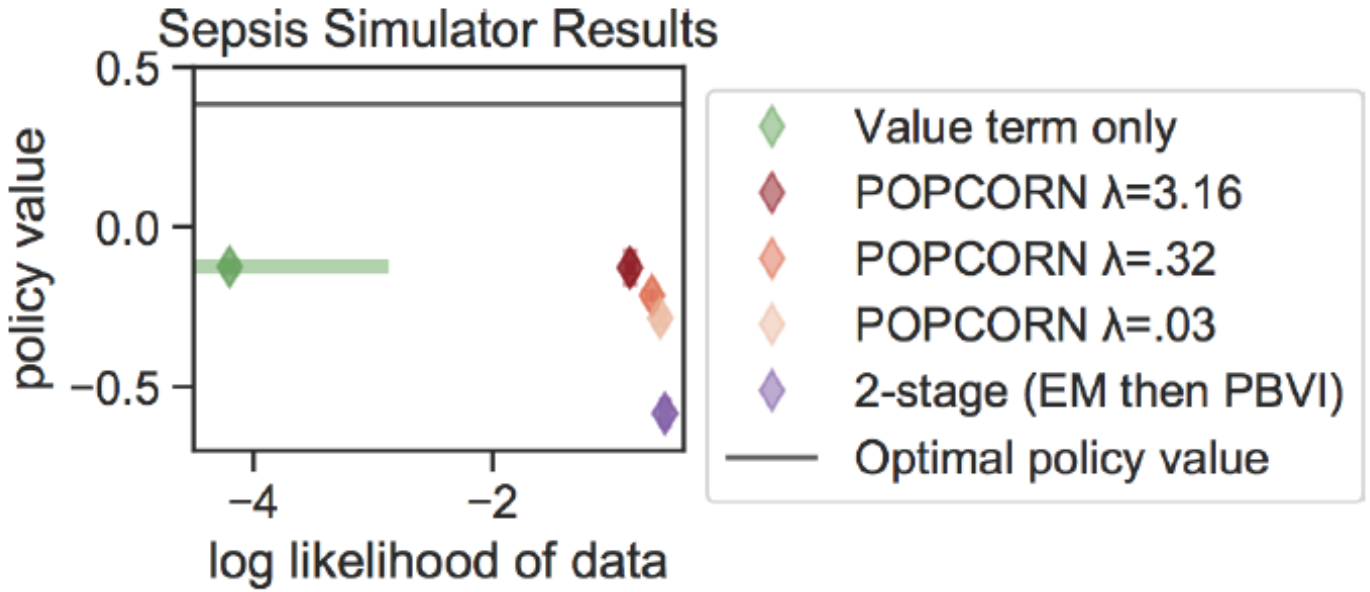
advantage of generative model



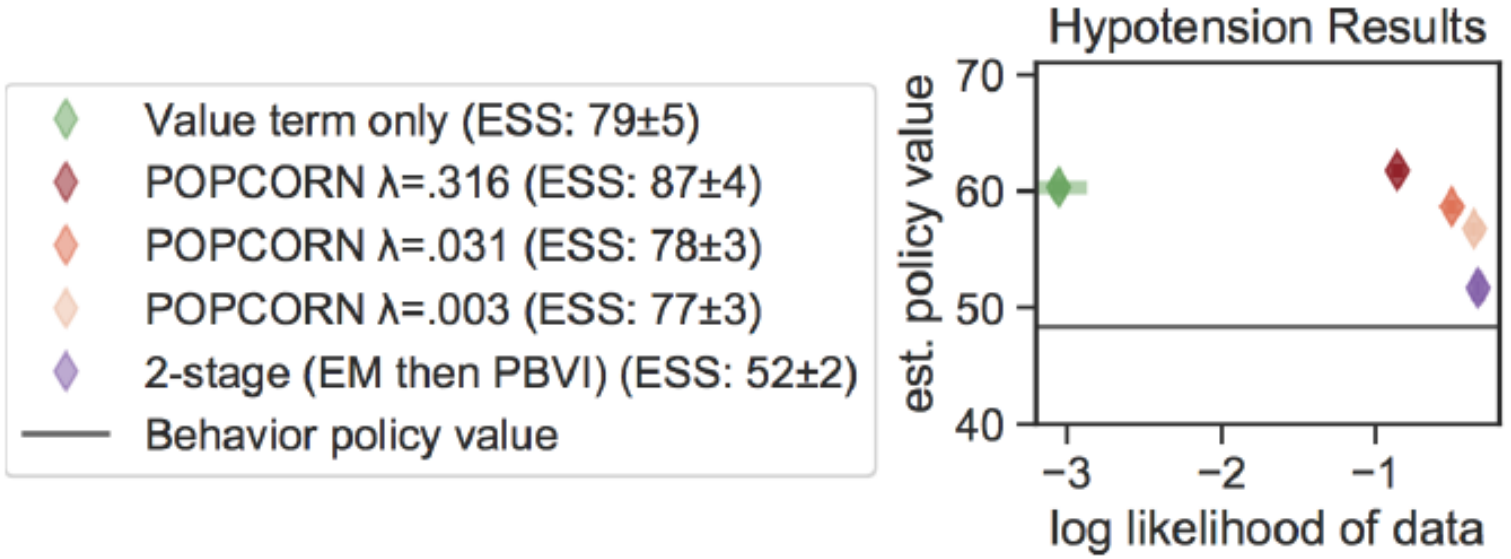
robust to misspecified model

Sepsis Simulator

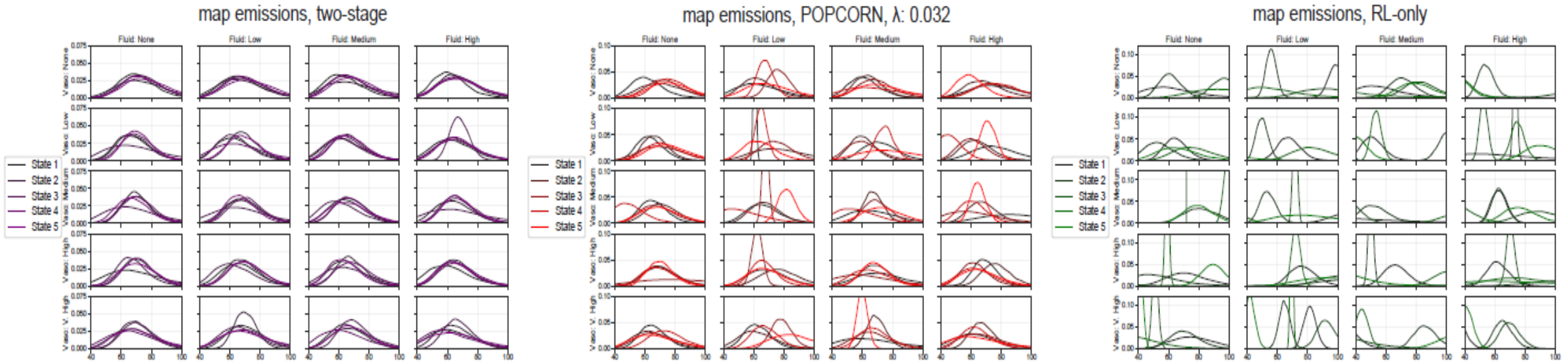
- Medically-motivated environment with known ground truth
- Results:



Real Data Application: Hypotension



Real Data Application: Hypotension



MAP: mean arterial pressure

Future directions

- Scaling to environments with more complex state structures
- Long-term temporal dependencies
- Investigating semi-supervised settings where not all sequences have rewards
- Ultimately become integrated into clinical decision support tools

References

- Komorowski, M., Celi, L.A., Badawi, O. *et al.* The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* **24,** 1716–1720 (2018). https://doi.org/10.1038/s41591-018-0213-5
- Raghu, Aniruddh, et al. "Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach." *arXiv preprint arXiv:1705.08422* (2017).
- Prasad, Niranjani, et al. "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units." *arXiv preprint arXiv:1704.06300* (2017).
- Ernst, Damien, et al. "Clinical data based optimal STI strategies for HIV: a reinforcement learning approach." *Proceedings of the 45th IEEE Conference on Decision and Control*. IEEE, 2006.
- Martín-Guerrero, José D., et al. "A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients." *Expert Systems with Applications* 36.6 (2009): 9737-9742.
- Hauskrecht, Milos, and Hamish Fraser. "Planning treatment of ischemic heart disease with partially observable Markov decision processes." *Artificial Intelligence in Medicine* 18.3 (2000): 221-244.
- Li, Luchen, Matthieu Komorowski, and Aldo A. Faisal. "The actor search tree critic (ASTC) for off-policy POMDP learning in medical decision making." *arXiv preprint arXiv:1805.11548* (2018).
- Oberst, Michael, and David Sontag. "Counterfactual off-policy evaluation with gumbel-max structural causal models." *arXiv preprint arXiv:1905.05824* (2019).
- Karkus, Peter, David Hsu, and Wee Sun Lee. "Qmdp-net: Deep learning for planning under partial observability." *Advances in Neural Information Processing Systems*. 2017.
- Igl, Maximilian, et al. "Deep variational reinforcement learning for POMDPs." *arXiv preprint arXiv:1806.02426* (2018).
- Pineau, Joelle, Geoff Gordon, and Sebastian Thrun. "Point-based value iteration: An anytime algorithm for POMDPs." *IJCAI*. Vol. 3. 2003.
- Rabiner, Lawrence R. "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE* 77.2 (1989): 257-286.