Off-Policy Interval Estimation with Lipschitz Value Iteration

Ziyang Tang[1], Yihao Feng[1], Na Zhang[2], Jian Peng[3], Qiang Liu[1]
[1]UT Austin, [2]Tsinghua University, [3]UIUC

Scenario

state s: patients physiological features.

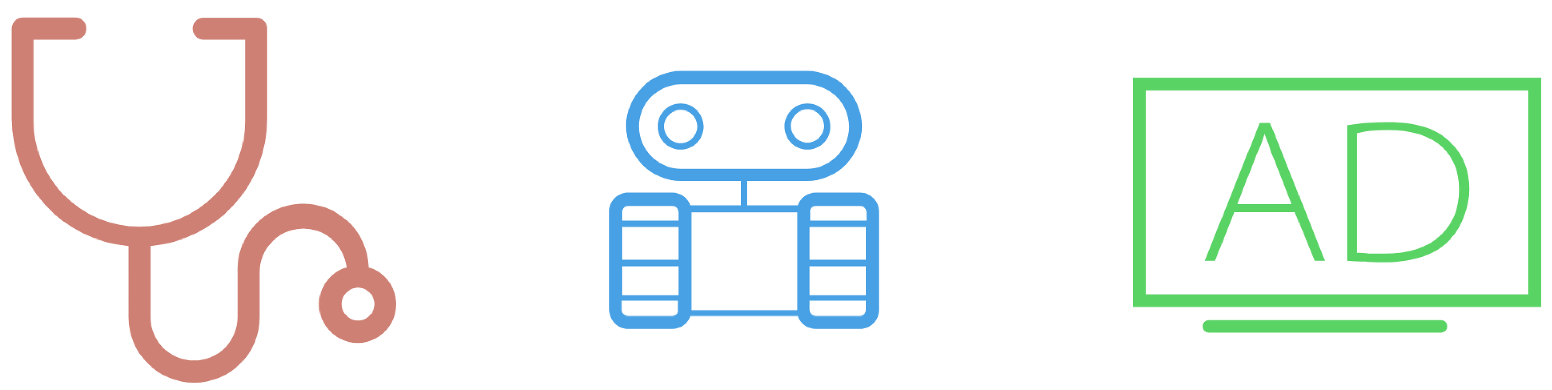
action a: medical action, e.g. take a medicine or not; how many doses.

reward r: patient condition; side effect.

next state s': patients physiological features at the next time step.

Off-Policy Evaluation(OPE)

- **Problem**: Evaluate a new policy π given *arbitrary* historical data.
- **Wide Application**: whenever evaluating new policies is costly or impossible, due to **high cost, risk, ethic or legal concerns**.



Healthcare **Robotic & Control** **Recommendation**

Hardness of Point Estimation

- Traditional OPE focus on point estimation, which can be arbitrarily bad due to:
  1. **High variance** for trajectories-based methods where variance grows exponentially with horizon length.
  2. **Bias** for optimization-based methods (e.g., value learning, model based method)
  3. **Small effective sample size** due to policy mismatch (distribution shift).
- In high-stakes scenarios, point estimation is not enough; need **confidence intervals** as well!!

Finite Samples Bellman Equations

- Formulate R^π with q-function:

$$R_\pi = \mathbb{E}_{s\sim\mu_0, a\sim\pi(\cdot|s)}[q_\pi(s,a)] := R_{\mu_0,\pi}[q_\pi].$$

- Notice that q_π is the **unique** fixed point of the Bellman equation:

$$q_\pi(x) = r(x) + \gamma\mathbb{E}_{s'=T(x), a'\sim\pi(\cdot|s')}[q_\pi(x')] := \mathcal{B}^\pi q_\pi(x),\ \forall x$$

  where x is short for state, action pair s, a.
- Only get access to **finite number** of the transition operator $\mathcal{B}^\pi$ which yields **not unique** fixed point solution.

Interval Estimation Frameworks

- Constraint on finite sample Bellman equation may lead to arbitrary large/small value on unseen region, need a model assumption $q_\pi \in \mathcal{F}$.
- **Optimization framework**:

$$\overline{R}_{\mathcal{F},\pi} = \sup_{q\in\mathcal{F}}\{R_{\mu_0,\pi}[q],\ \text{s.t. } q(x_i) = \mathcal{B}^\pi q(x_i),\ \forall i \in [n]\}.$$

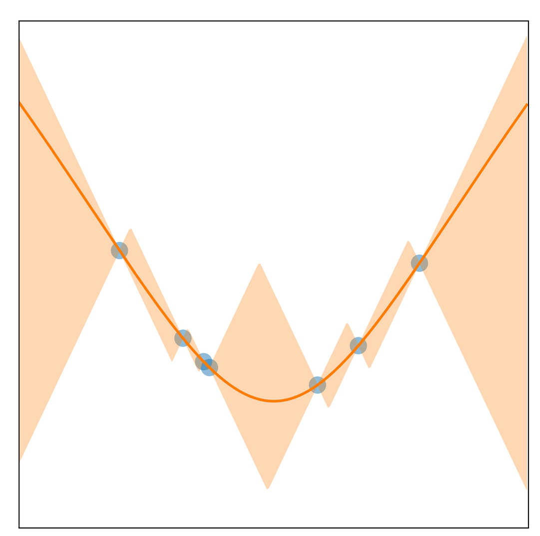
- **Simplest assumption**: *smoothness* assumption;
  Formally we consider the following bounded Lipschitz class:

$$\mathcal{F}_\eta = \{f : \|f\|_{Lip} \le \eta\},$$

  where $\|f\|_{Lip} := \sup_{x\neq x'} \frac{|f(x)-f(x')|}{d(x,x')}$.

A Lipschitz Regression Example

- Why is it possible to solve an infinite dimension optimization under finite sample constraints?



Consider a regression problem:

$$\overline{f}(x) = \sup_{f\in\mathcal{F}_\eta}\{f(x),\ \text{s.t. } f(x_i) = f_i,\ \forall i \in [n]\}$$

Closed form solution:

$$\overline{f}(x) = \min_{i\in[n]}\{f_i + \eta d(x, x_i)\}$$

- Similar for the lower bound: $\underline{f}(x) = \max_{i\in[n]}\{f_i - \eta d(x, x_i)\}$.

A Value Iteration Style Algorithm

Main Algorithm

Run the followings iteratively until convergence:

1. Plug in the last q_t as our new regression constraints

$$\overline{q}_{i,t+1} = \mathcal{B}^\pi\overline{q}_t(x_i).$$

2. Solve the new q_{t+1} as a Lipschitz regression problem:

$$\overline{q}_{t+1}(x) = \sup_{q\in\mathcal{F}_\eta}\{q(x),\ \text{s.t. } q(x_i) = \overline{q}_{i,t+1}\} = \min_{i\in[n]}\{\overline{q}_{i,t+1} + \eta d(x, x_i)\}.$$

Theoretical Properties of Lipschitz Value Iteration (Informal)

- **Monotonicity**, with a well-defined $\overline{q}_0$, we have:

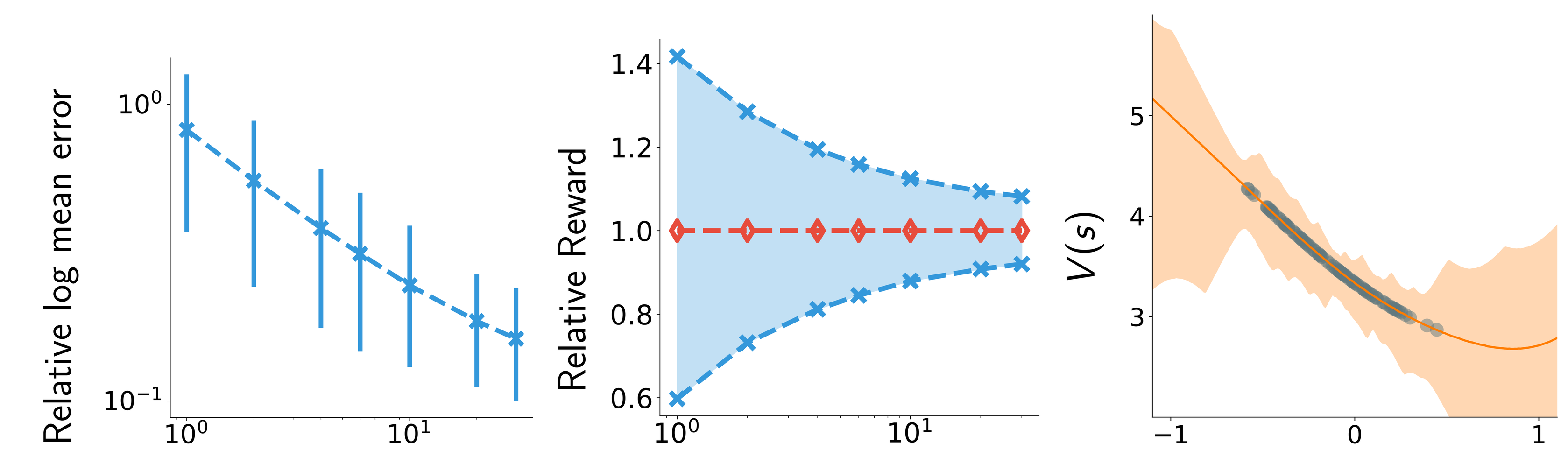
$$\overline{q}_t(x) \ge \overline{q}_{t+1}(x) \ge q_\pi(x),\ \forall x.$$

- **Linear Convergence:** $\overline{q}_t(x) - \overline{q}_\infty(x) = \mathcal{O}(\gamma^t)$.
- **Tightness of bounds:** $\overline{q}_t(x) - \underline{q}_t(x) = \mathcal{O}(\varepsilon_{X_n})$,
  where ε_{X_n} is the covering radius of data set $X_n = \{x_i\}_{i\in[n]}$, with:

$$\varepsilon_{X_n} = \sup_x \min_{i\in[n]}\{d(x, x_i)\}$$

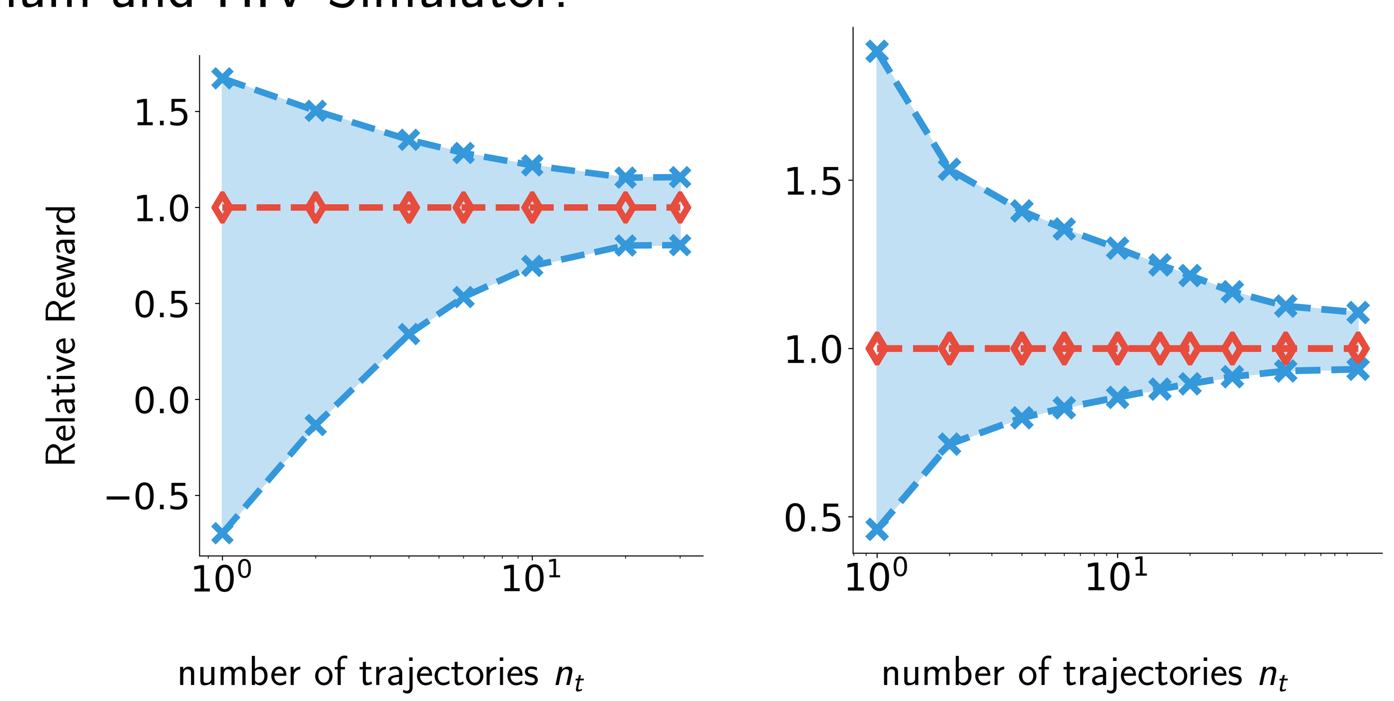
Experimental Results

- Synthetic Environment with a Known Value Function:



(a) number of trajectories n_t (b) number of trajectories n_t (c) State s

- Pendulum and HIV Simulator:



(a) Pendulum Environment (b) HIV Environment

Acknowledgment This work is supported in part by NSF CAREER #1846421, SenSE #2037267, and EAGER #2041327.