Replicability in Reinforcement Learning

Amin Karbasi, Grigoris Velegkas, Lin F. Yang, Felix Zhou

Yale, Google Research, UCLA

Replicability Crisis

nature

Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

Monya Baker

- Over 70% of researchers failed to replicate others' work
- Over 50% failed to replicate their own work!

- 2019 NeurIPS/ICLR Reproducibility Challenge (`github.com/reproducibility-challenge`)
- Ongoing ML Reproducibility Challenge (`papersithcode.com/rc2022`)

Goal: Mathematical Study of Replicability

- X data domain
- $\mathcal{D}$ distribution over X
- $S_1, S_2 \sim_{i.i.d.} \mathcal{D}^n$ size n datasets
- ξ random binary string

Definition (Replicable Algorithm) [Impagliazzo, Lei, Pitassi, Sorrell '22]
A randomized algorithm $\mathscr{A} : X^n \to Y$ is *ρ-replicable* if

$$\Pr_{S_1,S_2,\xi}[\mathscr{A}(S_1;\xi) = \mathscr{A}(S_2;\xi)] \geq 1-\rho.$$

Input: i.i.d. datasets, *shared* internal randomness

Goal: the output of the algorithm should be the same (w.h.p.)

Definition (TV Indistinguishable Algorithm) [Kalavasis, Karbasi, Moran, Velegkas '23]
A randomized algorithm $\mathscr{A} : X^n \to Y$ is *ρ-TV indistinguishable* if

$$\mathbb{E}_{S_1,S_2}[d_{\mathrm{TV}}(\mathscr{A}(S_1), \mathscr{A}(S_2))] \leq \rho.$$

Replicable Tabular Reinforcement Learning with Generative Model

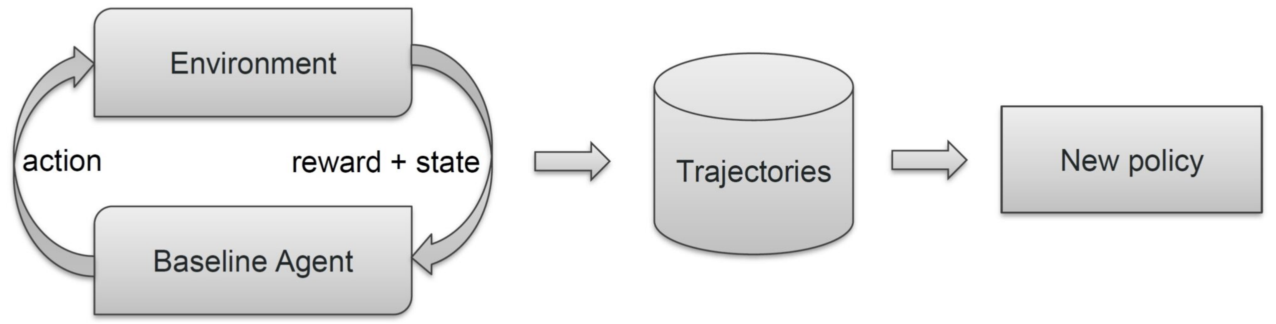
- **Given:** Generative model that gives samples of the *reward* and *transition* $r(s,a)$, $P(\cdot|s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
- **Want:** Output a policy $\pi : \mathcal{S} \to \mathcal{A}$
- Solve $\operatorname{argmax}_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0, P, \pi\right]$
- **And:** $\Pr_{S,S',\xi}\left[\pi_{S,\xi} = \pi'_{S',\xi}\right] \geq 1-\rho$

Main Result (Replicable Algorithms)

- Assume access to a generative model for the MDP.
- There is a ρ-replicable algorithm for the policy estimation problem such that:
- with probability at least $1-\delta$, outputs ε-approximate solution (additive).
- the algorithm has sample complexity

$$\tilde{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|^3\log(1/\delta)}{(1-\gamma)^5\varepsilon^2\rho^2}\right).$$

- The algorithm has polynomial running time (in the previous parameters).



Main Result (TV Indistinguishable Algorithms)

- Assume access to a generative model for the MDP.
- There is a ρ-TV indistinguishable algorithm for the policy estimation problem such that:
- with probability at least $1-\delta$, outputs ε-approximate solution (additive).
- the algorithm has sample complexity

$$\tilde{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|^2\log(1/\delta)}{(1-\gamma)^5\varepsilon^2\rho^2}\right).$$

- The algorithm has polynomial running time (in the previous parameters).

Remark

We can transform the TV indistinguishable algorithm to a replicable one, but we need time $\exp(|\mathcal{S}| \cdot |\mathcal{A}|)$.

Overview of Techiques

0. Get enough samples to estimate a Q-function $\widehat{Q}$ such that $||\widehat{Q} - Q^\star||_\infty$ is sufficiently small
   - Many techniques from the RL literature
1. Across the two executions we have that $||\widehat{Q}_1 - \widehat{Q}_2||_\infty$ is sufficiently small
   - Replicable Algorithm: Use randomized rounding scheme from [Impagliazzo, Lei, Pitassi, Sorrell '22] to get $\widehat{Q}_1 = \widehat{Q}_2$
   - TV Indistinguishable Algorithm: Novel technique based on the Gaussian mechanism from the DP literature (coulde be of independent interest)
2. From Q-function approximation to policy estimation:
   - Replicable Algorithm: Use greedy policy w.r.t. the estimated Q-function
   - TV Indistinguishable Algorithm: Use greedy policy w.r.t. the estimated Q-function + data-processing inequality
3. Lower bound for a class of algorithms

Can we get sample complexity linear in $|\mathcal{S}| \cdot |\mathcal{A}|$?

- Replicability and TV indistinguishability impose a discrete metric on the policy space
- **Idea:** Consider a more fine-grained notion of distance over policies
  - Treat policies as probability distributions over actions

Approximate Replicability

- $\mathcal{S}$ state space
- $\mathcal{G}$ generative model of the MDP
- κ dissimilarity measure of distributions (e.g., KL divergence)
- $S_1, S_2 \sim_{i.i.d.} \mathcal{G}$ i.i.d. samples from the generative model
- ξ random binary string

Definition (Approximately Replicable Policy Estimator)
A randomized algorithm $\mathcal{A}$ is (ρ_1, ρ_2)-*approximately replicable* if

$$\Pr_{S_1,S_2,\xi}\left[\max_{s\in\mathcal{S}} \kappa(\pi_1(s), \pi_2(s)) \geq \rho_1\right] \leq \rho_2,$$

where π_1, π_2 is the output of the algorithm on $(S_1;\xi), (S_2;\xi)$, respectively.

Main Result (Approximately Replicable Algorithms)

- Assume access to a generative model for the MDP.
- There is a (ρ_1, ρ_2)-replicable algorithm for the policy estimation problem such that:
- with probability at least $1-\delta$, outputs ε-approximate solution (additive).
- the algorithm has sample complexity

$$\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\log(1/(\delta\cdot\rho_2))}{(1-\gamma)^5\varepsilon^2\rho_1^2}\right).$$

- The algorithm has polynomial running time (in the previous parameters).

Future Work

- Improve upper bounds for replicable algorithms
- Establish lower bounds for TV indistinguishable algorithms
- Improve dependence on γ
- Study the general function approximation setting
- Study the online setting

felix.zhou@yale.edu