Adversarial Learning of Distributional Reinforcement Learning

Yang Sui[1] Yukun Huang[1] Hongtu Zhu[2] Fan Zhou[1]

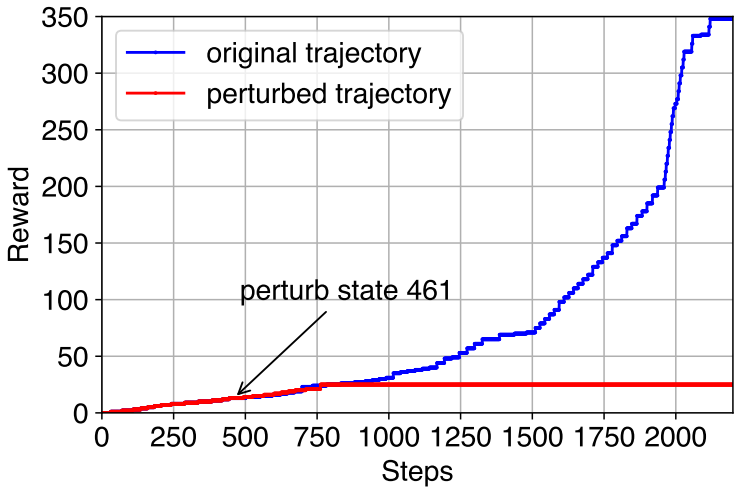
[1]Shanghai University of Finance and Economics

[2]The University of North Carolina at Chapel Hill

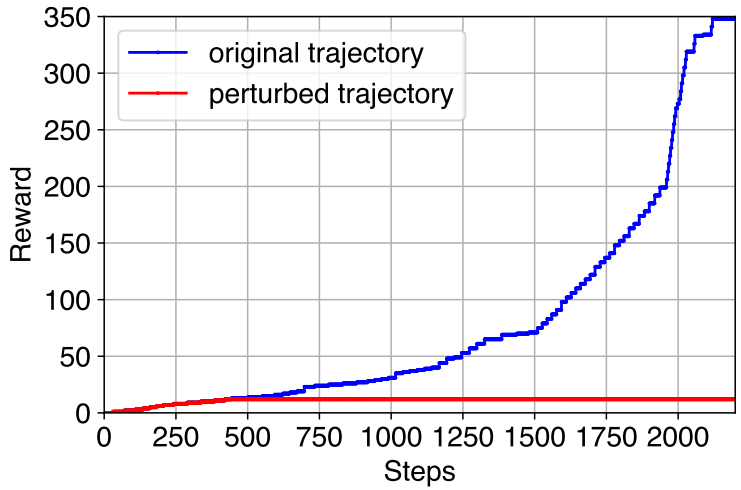
Motivation: Ride-Sharing

- The order-dispatching policy is trained offline which is then applied to real world
- Although offline and online environments may be almost the same, these policies can perform poorly in real world

Motivation: Empirical Studies



perturb state



perturb parameter

- Small perturbation imposed on certain state can significantly change the trajectory afterwards
- Similar result after perturbing some parameter in the policy network

Our Contributions

- Question: **How to quantify the effects of these variations?**
- Prior Work: **distribution shift between online and offline data**
- Our Contributions:
  - analyse why a trained policy fail when applied to new but similar environment
  - construct adversarial learning to evaluate sensitivity of RL components
  - our method is efficient for detecting DRL, theoretically and empirically

Method: Perturbation Mainfold

- **DRL Setting:** Given state $\boldsymbol{s}$, action $\boldsymbol{a}$ and network parameter $\boldsymbol{\theta}$, the distribution of $\boldsymbol{z}$ is $P(\boldsymbol{z}|\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta})$.

  The perturbed model is $P(\boldsymbol{z}|\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\omega})$ after imposing a perturbation $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T$ on either $\boldsymbol{s}$ or $\boldsymbol{\theta}$

- **Perturbation Manifold:** $M = \{P(\boldsymbol{z}|\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\omega}) : \boldsymbol{\omega} \in \Omega\}$. The vector space of M at $\boldsymbol{\omega}$ is spanned by p functions $\{\partial_i \ell(\boldsymbol{\omega}|\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta})\}_{i=1}^{p}$, where $\partial_i = \partial/\partial\omega_i$ and $\ell(\boldsymbol{\omega}|\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta}) = \log P(\boldsymbol{z}|\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\omega})$ [1, 2]

- **Metric Tensor:** $\boldsymbol{G}(\boldsymbol{\omega}) \in \mathbb{R}^{p \times p}$ of $\boldsymbol{\omega}$ is defined as

$$g_{ij}(\boldsymbol{\omega}) = \boldsymbol{E}_{\boldsymbol{\omega}} \left[\partial_i \ell(\boldsymbol{\omega}|\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta}) \partial_j \ell(\boldsymbol{\omega}|\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta})\right]$$

Influence measure

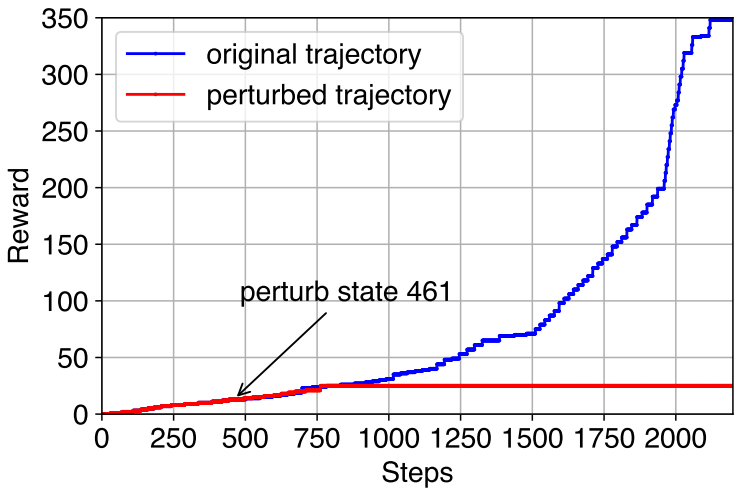
Let $f(\boldsymbol{\omega}) : \mathbb{R}^p \to \mathbb{R}^1$ be the objective function, we can then define the first-order local influence measure $(\mathbf{FI})$ of $f(\boldsymbol{\omega})$ at $\boldsymbol{\omega}_0$ as

$$\mathbf{FI}_{\boldsymbol{\omega}}(\boldsymbol{\omega}_0) = \nabla^T_{f(\boldsymbol{\omega}_0)} \boldsymbol{G}^{-1}(\boldsymbol{\omega}_0) \nabla_{f(\boldsymbol{\omega}_0)}$$

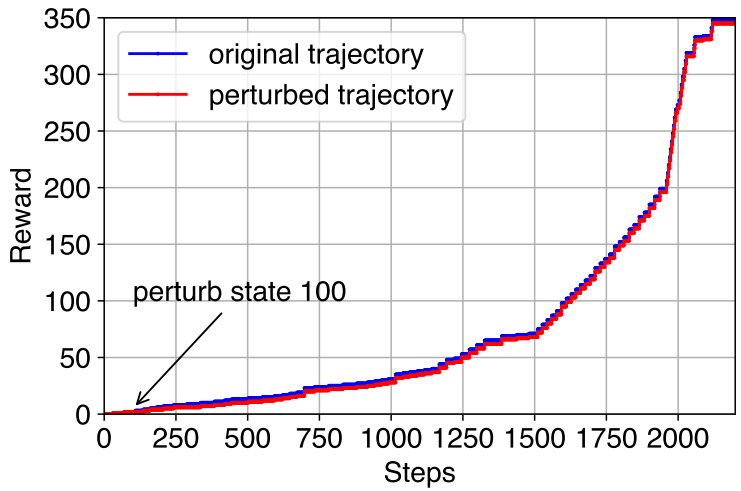
- High $\mathbf{FI}$ represents that $f(\boldsymbol{\omega})$ is more vulnerable to $\boldsymbol{\omega}$
- $\mathbf{FI}$ possesses an intrinsic property that is free of the constraints imposed by $\boldsymbol{\omega}$
- $\mathbf{FI}$ is invariant to any reparameterization of $\boldsymbol{\omega}$

Task (i): Detection of fragile states

Comparison of trajectories for states (Q-values as objective function)



state 461 with high $\mathbf{FI}$:
significant difference

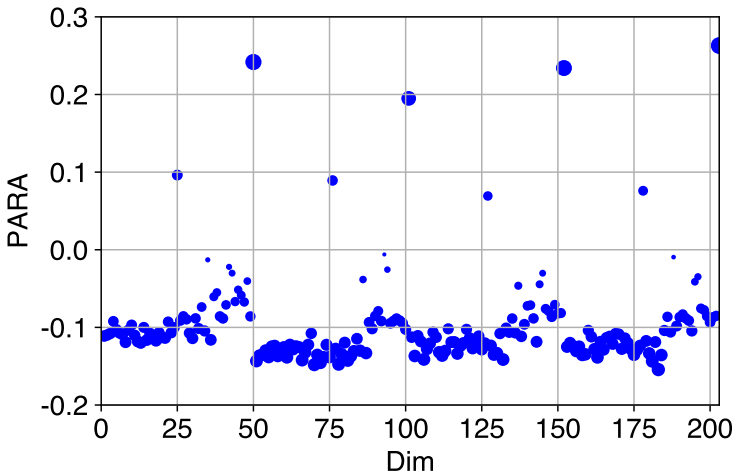


state 100 with low $\mathbf{FI}$:
negligible difference

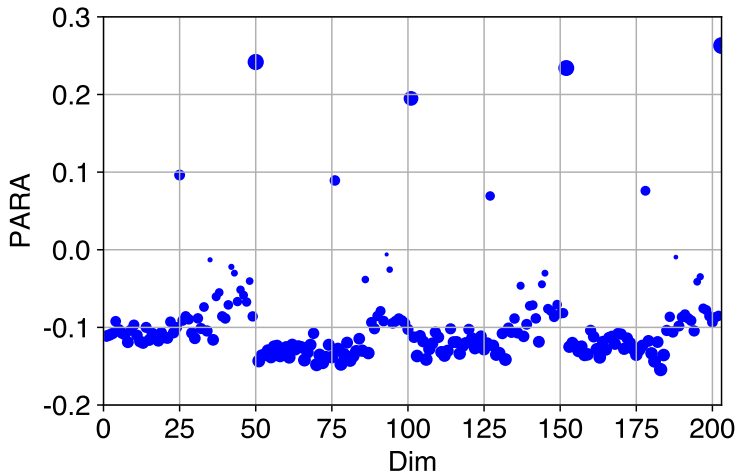
$\mathbf{FI}$ can detect potentially vulnerable states!

Task (ii): Adversarial learning of policy network

Parameters in the DENSE2 BIAS layer and their corresponding $\mathbf{FI}$s



parameters

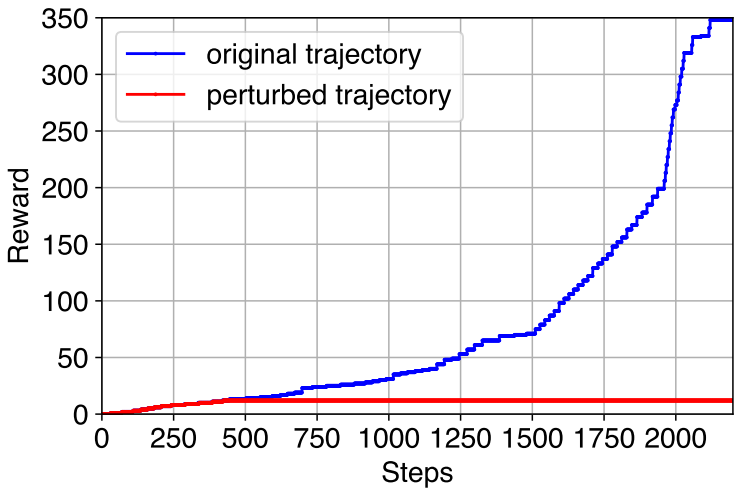


$\mathbf{FI}$**s**

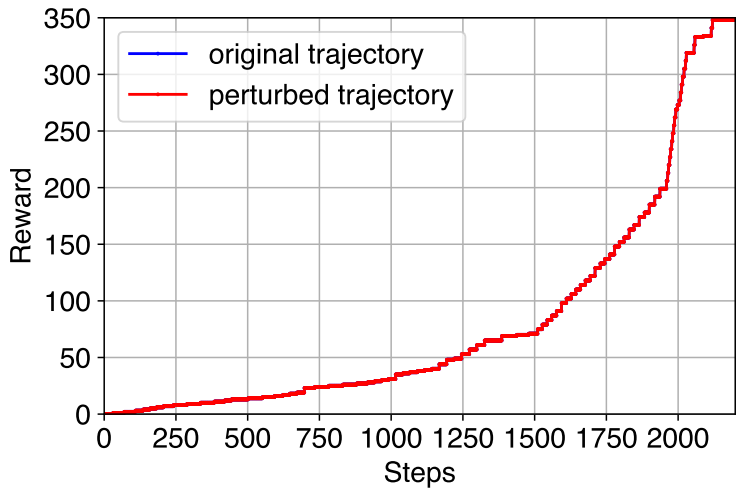
Only 8 positive parameters $\Longleftrightarrow$ 8 parameters with high $\mathbf{FI}$ detected

Task (ii): Adversarial learning of policy network

Comparison of trajectories for parameters (Q-values as objective function)



204th parameter with high $\mathbf{FI}$:
significant difference



152nd parameter with low $\mathbf{FI}$:
no difference

$\mathbf{FI}$ can pinpoint sensitive policy parameters!

References

[1] H. Zhu, J. G. Ibrahim, S. Lee, and H. Zhang. Perturbation selection and influence measures in local influence analysis. *The Annals of Statistics*, 35(6):2565–2588, 2007.

[2] H. Zhu, J. G. Ibrahim, and N. Tang. Bayesian influence analysis: a geometric approach. *Biometrika*, 98(2):307–323, 2011.