Reinforcement Learning Gradients as Vitamin for Online Finetuning Decision Transformers

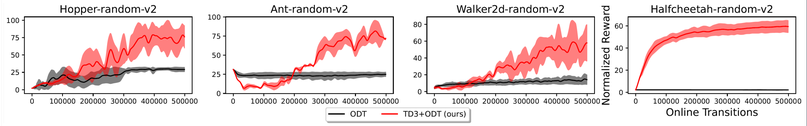
Kai Yan, Alexander G. Schwing, Yu-Xiong Wang

NeurIPS 2024 Spotlight

arXiv: 2410.24108

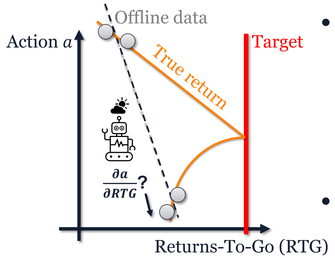
Motivation

Goal: analyze and address Online Decision Transformer [1] (ODT)'s struggle during online finetuning upon pretraining with low-return data

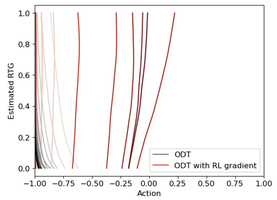


Online finetuning after pretraining on MuJoCo random dataset; expert-level reward is 100

Case Study: Single-State MDP



- When target Returns-To-Go (RTG) is close, ODT can improve via **conditioning on high target RTG** and entropy terms
- But this is **misleading** when pretrain on data where target RTG is too **out-of-distribution!**
- Worse still, ODT cannot improve in local action space because **ODT gives $\frac{\partial a}{\partial RTG}$, but we need $\frac{\partial RTG}{\partial a}$!**



ODT struggles to improve

Theoretical Analysis

How does out-of-distribution affect performance?

We resort to Brandfonbrener et al.'s [2] **tight** bound:

$$\mathbb{E}[RTG_{\text{target}}] - \mathbb{E}[RTG_{\text{true}}] \leq O\left(\frac{1}{P_{\text{data}}(RTG_{true} = RTG_{\text{target}})}\right)$$

Linear with respect to (w.r.t.) RTG_{target}

Decrease to fit in tight bound with growing RTG_{target}

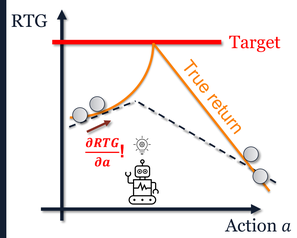
Grows **superlinearly** w.r.t. RTG_{target}!

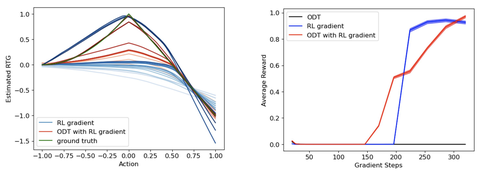
*informal, with unique initial state; check paper for rigorous formula

Interesting fact: ODT's policy improvement resembles that of AWAC but requires better global RTG property. Check our paper for details!

Practical Solution

- ODT needs gradient of $\frac{\partial RTG}{\partial a}$ for policy improvement in local action space – exactly what RL does!
- We train a MLP critic, and add to the ODT loss a down-weighted RL actor loss, where the decision transformer serves as the actor
- We find TD3 to be the most effective RL loss choice

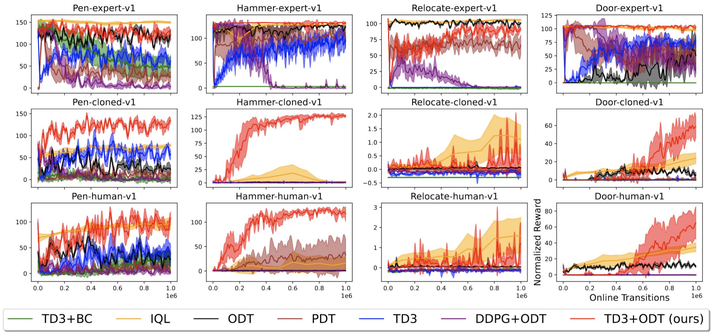




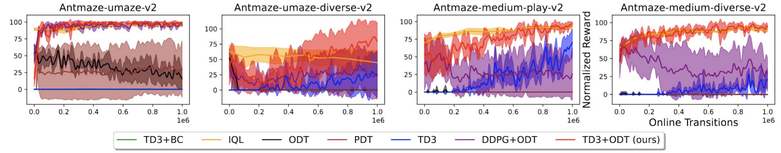
RL gradient helps to discover hidden reward peak

Results

We test our solution on adroit, MuJoCo, antmaze and maze2d with 30+ different datasets



Reward curves on adroit environments; ours illustrated in red (higher is better)



Reward curves on antmaze environments; ours illustrated in red (higher is better)

Check our paper for 10+ ablations!

Conclusion

- **Our key contribution:**
  - Study a largely underexplored problem of online finetuning decision transformers with low-return pretraining data
  - Give theoretical analysis and simple but effective solution
  - Conduct detailed and extensive empirical evaluations
- **Limitation:**
  - ~20% extra computational cost
  - Did not test on image-based environments

Key Papers

[1] Q. Zheng et al. Online Decision Transformer. In ICML, 2022.
[2] D. Brandfonbrener et al. When Does Return-Conditioned Supervised Learning Work for Offline Reinforcement Learning? In NeurIPS, 2022.