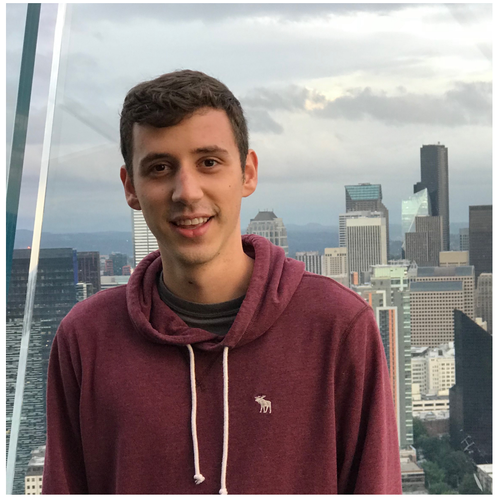
Multi-Stage Episodic Control for Strategic Exploration in Text Games

ICLR 2022 (Spotlight)

Jens Tuyls[1], Shunyu Yao[1], Sham Kakade[2], Karthik Narasimhan[1]

[1]Princeton University, [2]Harvard University



Text Games - Motivation

Partially Observable Markov Decision Process (POMDP)

- Player receives observations and rewards, issues actions
- States and model are hidden

Key challenges

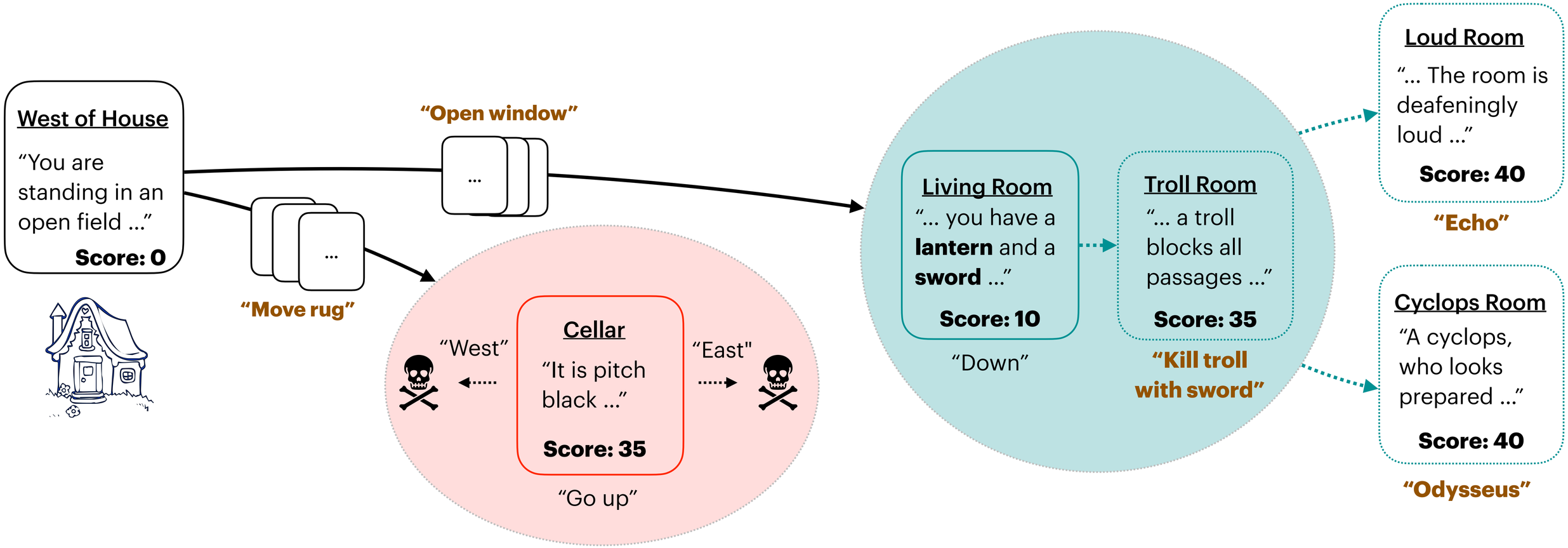
1. Sparse rewards
2. Large, dynamic action spaces (e.g. “Kill troll with sword”)

eXploit-Then-eXplore (XTX)

1. **Exploitation** phase: keeps track of a **global** “policy cover” of the game space
    A. Similar to how a human player returns back to promising parts of the game
    B. Allows quick learning from **sparse rewards**! ✓
2. **Exploration** phase: performs strategic **local** exploration based on past knowledge and action uncertainty estimates
    1. Once at the frontier, human player cleverly explores the frontier locally
    2. Allows to strategically explore the **large, dynamic action space** ✓

eXploit-Then-eXplore (XTX)

1. **Exploitation** phase: keeps track of a **global** “policy cover” of the game space
2. **Exploration** phase: performs strategic **local** exploration



XTX as a policy mixture

Action a, context c, observation o

$$\pi_\lambda(a \mid c, o) = \lambda \pi_{\text{inv-dy}}(a \mid o) + (1 - \lambda)\pi_{\text{il}}(a \mid c)$$

Exploration policy based on Q values and Inverse Dynamics (inv-dy) bonuses

Exploitation policy that returns the agent to the game frontier with Imitation Learning (il) on online trajectories

λ determines **trade-off**

(Dynamically changed in episode)

eXploit-Then-eXplore - Episodic Rollouts

For every episode E:

  For every time t:

    If **PHASE 1**: $\lambda = \frac{1}{2T}$

    Elif **PHASE 2**: $\lambda = 1$

    Sample action from π_λ

    Update $\pi_{\text{inv-dy}}$ with TD loss

  If n episodes passed:

    Update π_{il} on new online trajectories $\mathcal{B}$

$$\pi_\lambda = \lambda \pi_{\text{inv-dy}} + (1 - \lambda)\pi_{\text{il}}$$

Evaluation & Baselines

Evaluate on 12 human-created games from **Jericho** (Hausknecht et al., 2020)

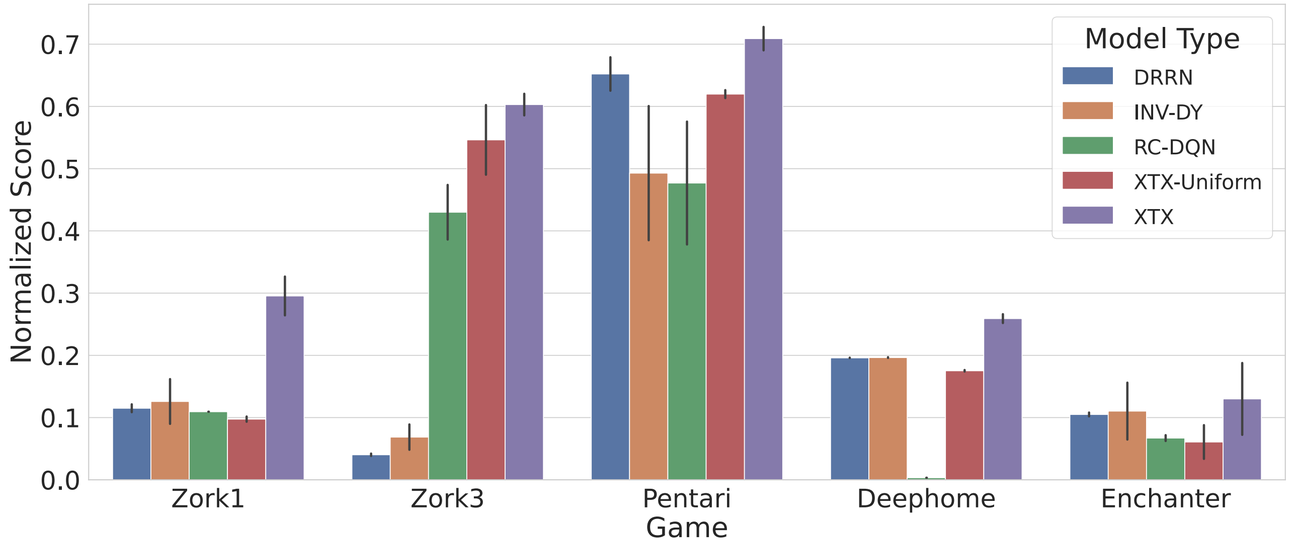
- Variety of challenges such as darkness, inventory management, etc.
- Deterministic and **stochastic setting** (observation + transition randomness)

Baselines

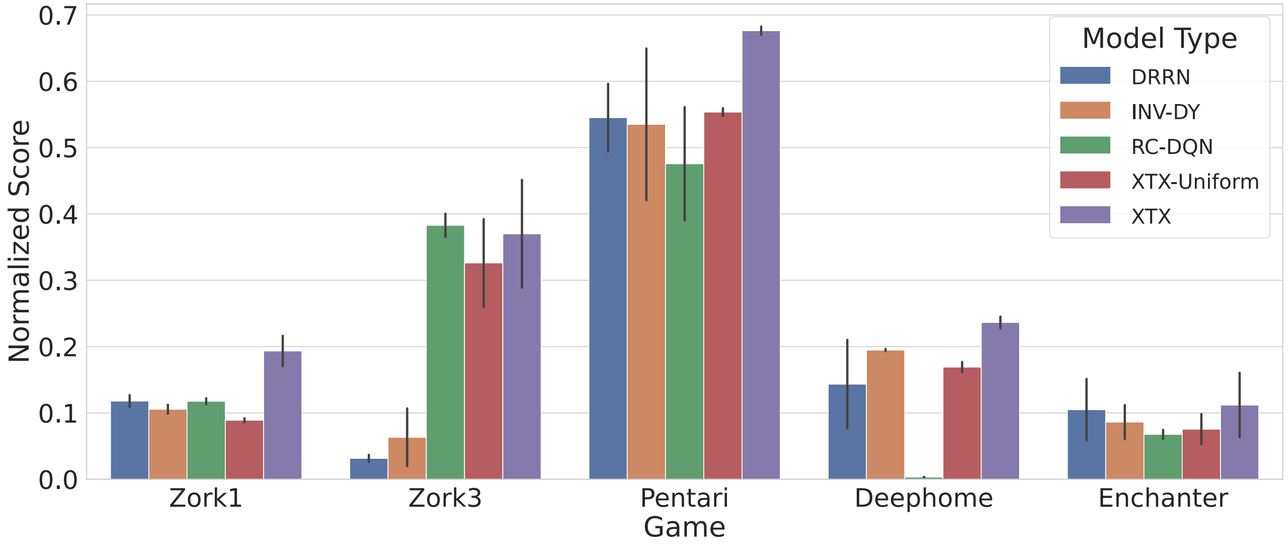
- DRRN (He et al., 2016)
- INV-DY (Yao et al., 2021)
- RC-DQN (Guo et al., 2020)
- XTX-Uniform (~Go-Explore (Ecoffet et al., 2021))

Results - Avg. Scores

Normalized score = raw games score/max score



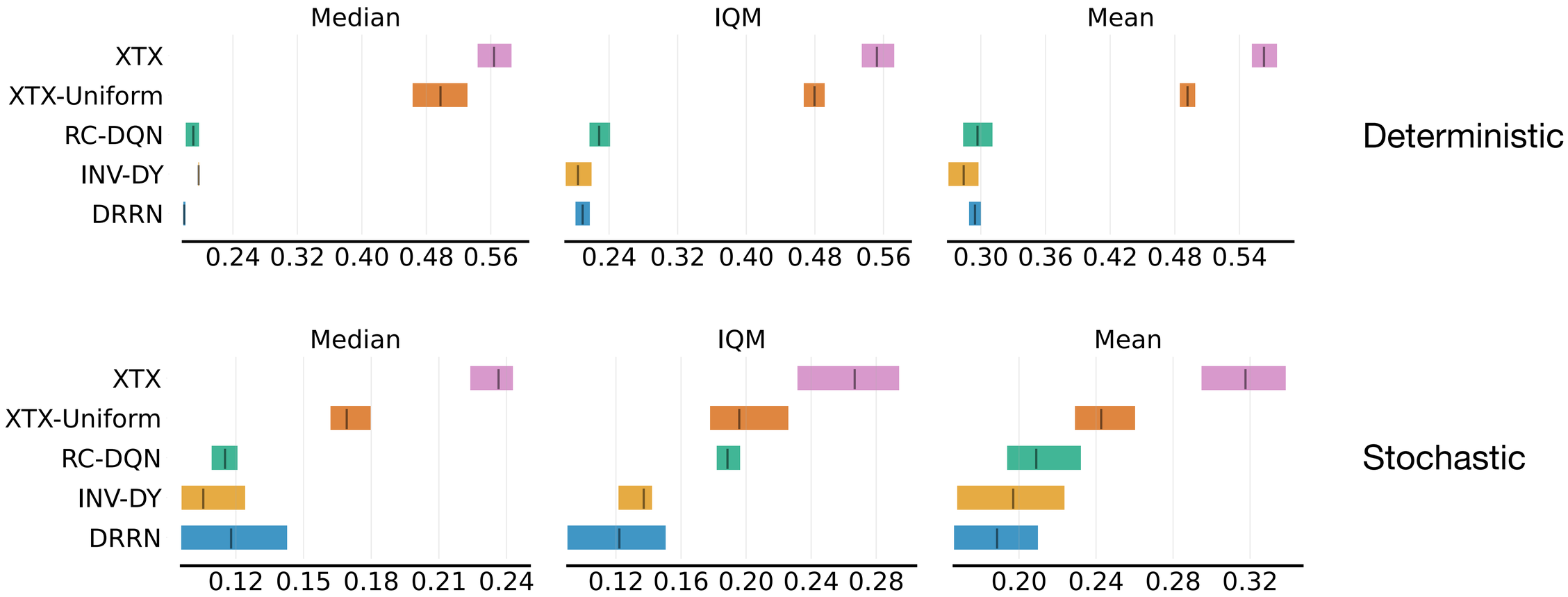
Deterministic



Stochastic

Results - Summary

Normalized score = raw games score/max score



(Agarwal et al., 2021)

Conclusion

We propose **eXploit**-Then-**eXplore** (XTX) to solve key challenges in text games:

(1) sparse rewards and (2) large, dynamic action spaces

XTX

1. **Explicitly separates** episodic rollouts into exploitation and exploration
2. Keeps track of a **global policy cover** with **strategic local exploration** at the frontier

✓ SOTA on 12/12 deterministic and 4/5 stochastic games in Jericho

✓ Overall significantly outperforms prior methods

✓ More than 2x improvement on famous Zork1!

Thank you.