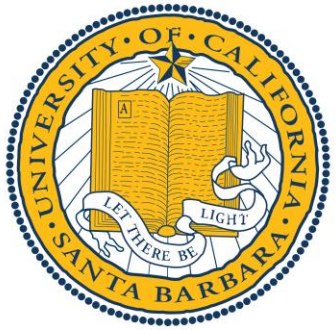
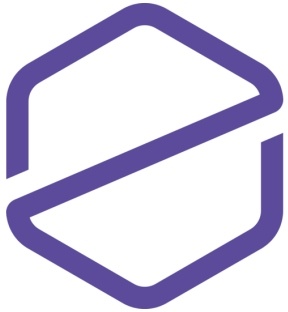
What makes math problems hard for reinforcement learning: a case study

Ali Shehper, Anibal Medina-Mardones, Lucas Fagan, Bartłomiej Lewandowski, Angus Gruen, Yang Qiu, Piotr Kucharski, Zhenghan Wang, Sergei Gukov

Western





State of Reinforcement Learning

- Success in Board and Video Games
  - Chess, Shogi, Go, Poker
  - Atari, Dota, StarCraft
- Math is the next playground
  - Theorem-Proving
  - Research Problems as RL Environments

Research-Level Math as RL Playground

- Andrews-Curtis Conjecture (1965)
  - Long Horizons
  - Sparse Rewards
  - A non-uniform distribution of hardness
- Existing Algorithms: Success and Limitations
- New Mathematical Results
- Propose New Algorithms

Andrews-Curtis Conjecture

- State Space: Presentations $\langle x, y \mid r_1, r_2 \rangle$ of the trivial group
- Action Space: Andrews-Curtis (AC) moves

1. Substitute some r_i by $r_i r_j$ for $i \neq j$.
2. Replace some r_i by r_i^{-1}.
3. Change some r_i to $x_j^{\pm 1} r_i x_j^{\mp 1}$.

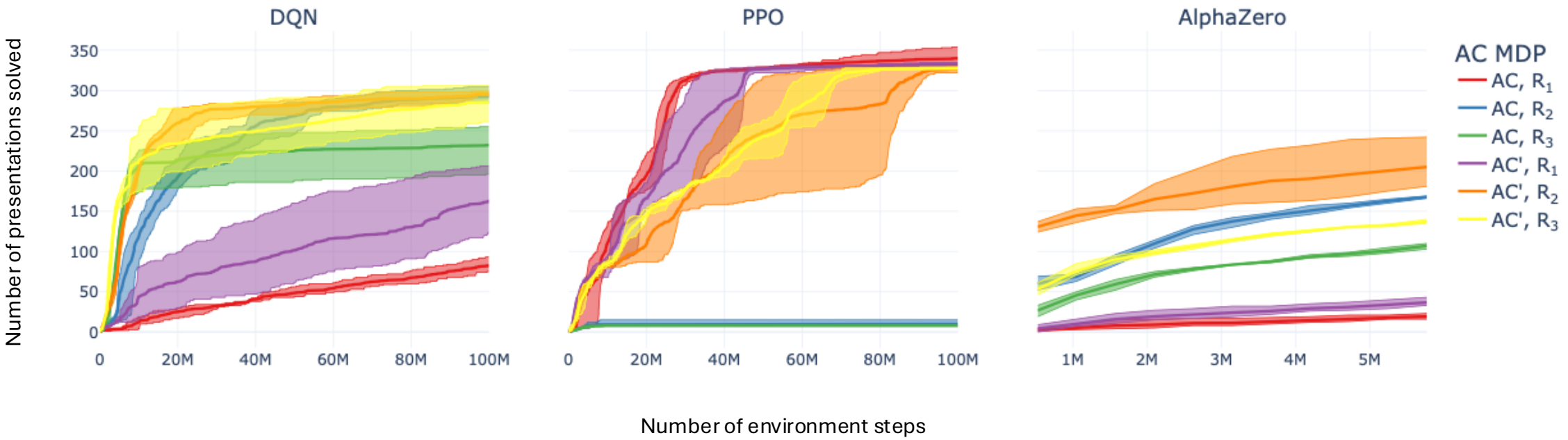
- Goal State: $\langle x, y \mid x, y \rangle$
- Rewards are sparse

Examples of Interest

- Open potential counterexamples
  - Akbulut-Kirby Series (1985): $\mathbf{AK}(n) = \langle x, y \mid x^n = y^{n+1},\ xyx = yxy \rangle.$
  - Miller-Schupp Series (1999): $\mathbf{MS}(n, w) = \langle x, y \mid x^{-1}y^n x = y^{n+1},\ x = w \rangle.$

- Solved with super-exponentially long solutions
  - Bridson, Lishak (2015): solutions of length 10^{10000}

DQN vs PPO vs AlphaZero



New Mathematical Results

Theorem A. The following infinite subfamilies of Miller–Schupp presentations are AC-trivial:

1. $\mathrm{MS}(1, w)$ for all w.
2. $\mathrm{MS}(n, y^{-1}xyx^{-1})$ for all n.
3. $\mathrm{MS}(2, y^{-k}x^{-1}yxy)$ for all k.

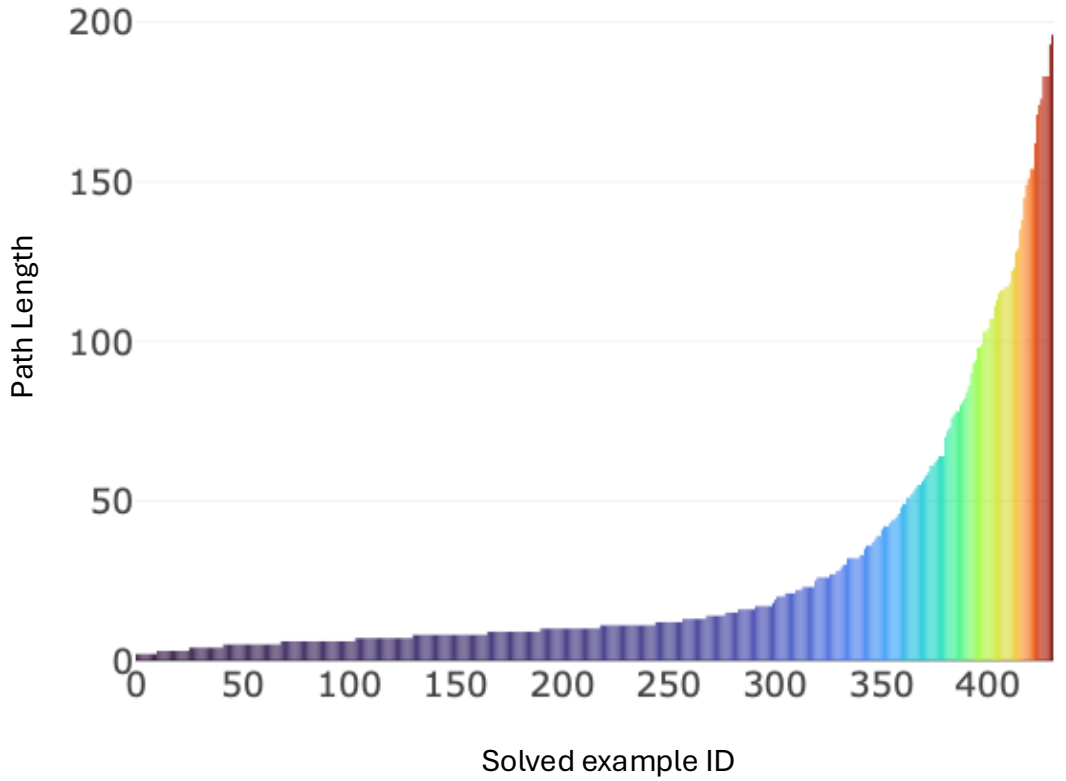
Theorem B. For every $n \geq 2$, $\mathrm{AK}(n)$ is AC-equivalent to the presentation

$$\langle x, y \mid x^{-1}yx = xyx^{-1}y \ , \ xy^{n-1}x = yxy \rangle,$$

of length $n + 11$. This gives a reduction in length of $AK(n)$ for all $n \geq 5$.

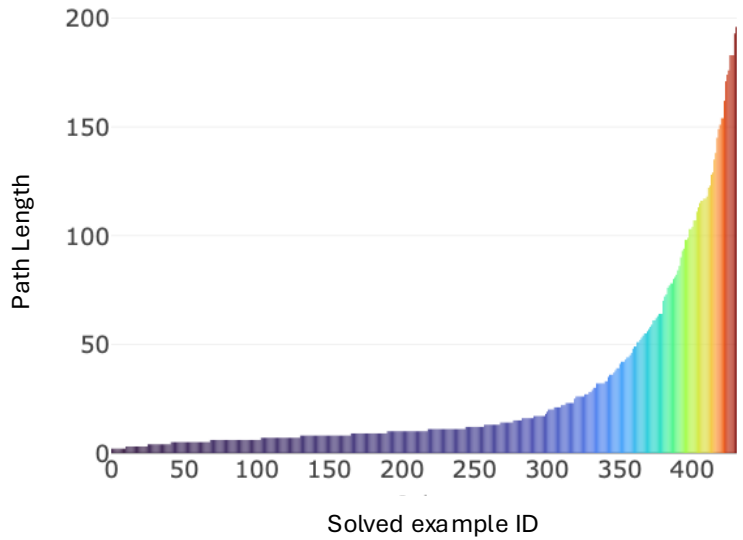
Limitations of Existing RL Algorithms

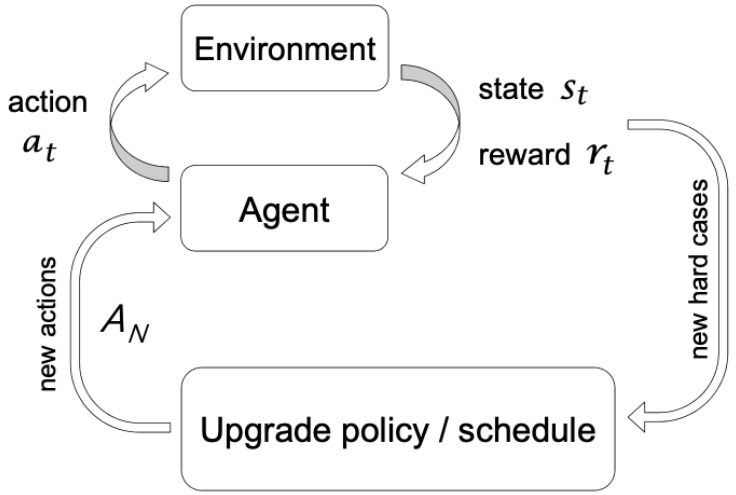
- Struggle with discovery of long paths
- Super-exponentially long paths in polynomial time?



New Reinforcement Learning Algorithms?

- Proposed Solution
  - Supermoves
  - Adaptive Action Spaces
  - Use hardness of states (such as path lengths) to select new actions (supermoves) during training
- Preliminary Experiments show success





Thank you

San Diego Convention Center

Exhibit Hall C,D,E

Thu 4 Dec 4:30 p.m. PST — 7:30 p.m. PST