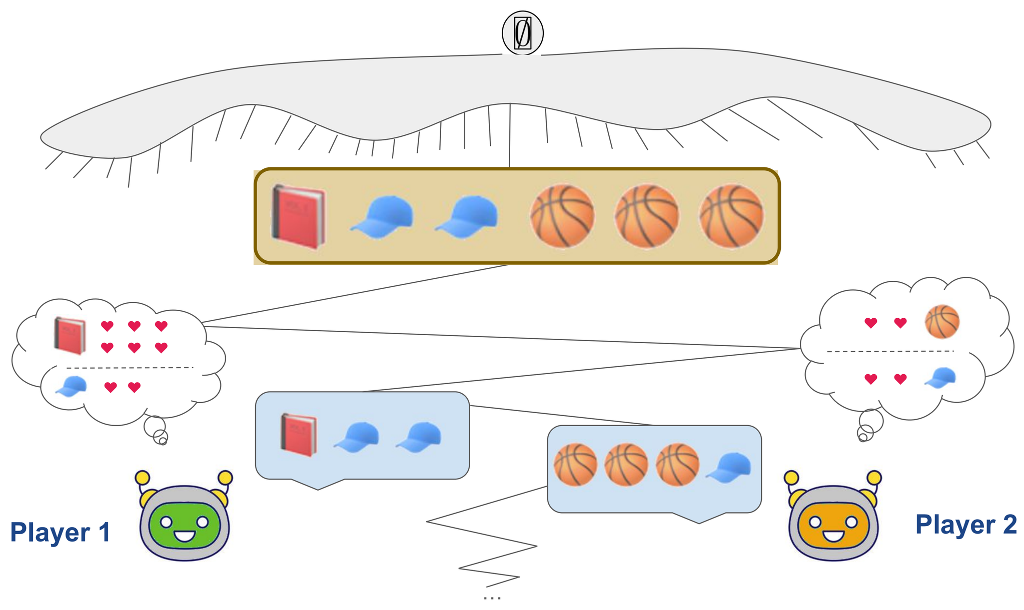
Combining Deep RL and Search with Generative Models for Game-Theoretic Opponent Modeling

Zun Li[1,2], Marc Lanctot[1], Kevin McKee[1], Luke Marris[1], Ian Gemp[1], Daniel Hennes[1], Paul Muller[1], Kate Larson[1,3], Yoram Bachrach[1], Michael P. Wellman[2]
[1]DeepMind, [2]University of Michigan, Ann Arbor, [3]University of Waterloo

Motivated Example: Deal-Or-No-Deal [1] Alternating Bilateral Negotiation Games

There are several challenges in solving large general-sum imperfect information games:

- Large state space.
- The large imperfect information (i.e., private values of the opponents) in the games may prohibit efficient planning.
- The equilibrium selection problem, i.e., how to train agents that can generalize well against unknown opponents at test-time. Naive self-play methods may overfit to the partner at training time.



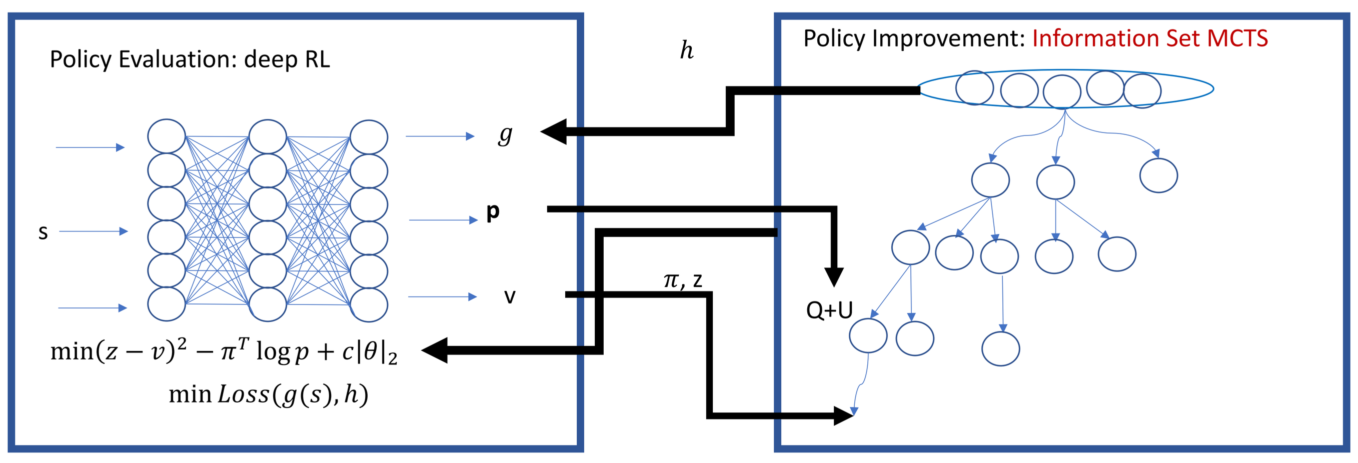
Extending AlphaZero to Imperfect Information

AlphaZero-styled search iteratively:

- Train a policy-and-value-network (PVN) using trajectories generated by MCTS
- Use the PVN to guide the search procedure and produce more quality data

To couple with imperfect information, we:

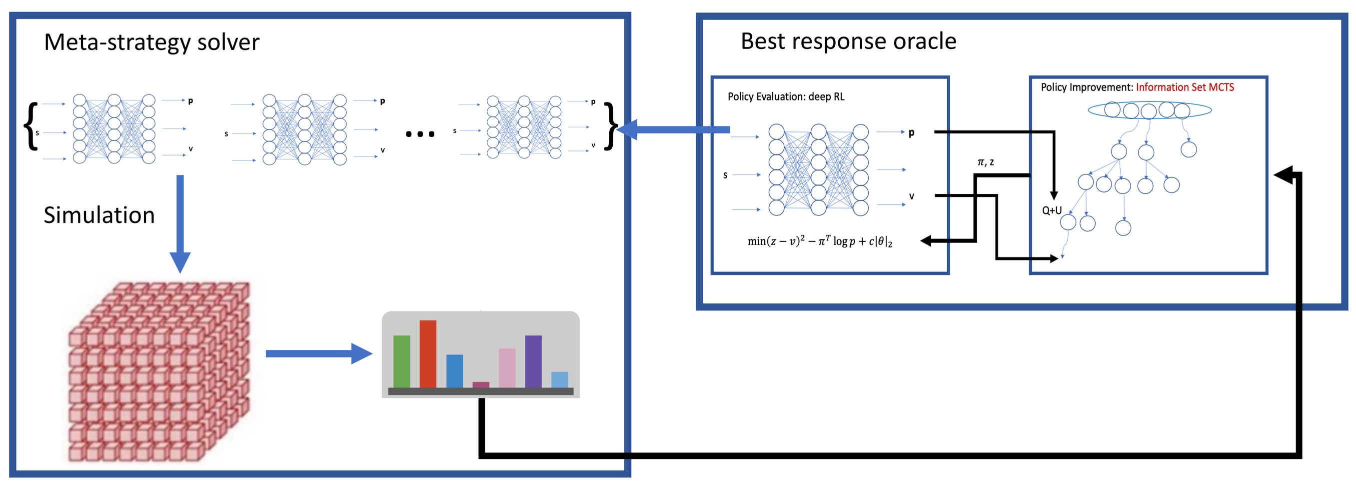
- Use information-set MCTS (IS-MCTS) as the search procedure
- At the root of the search tree, add a deep generative model to represent belief state
- Train the generative model using the (infostate, state) in the RL trajectory



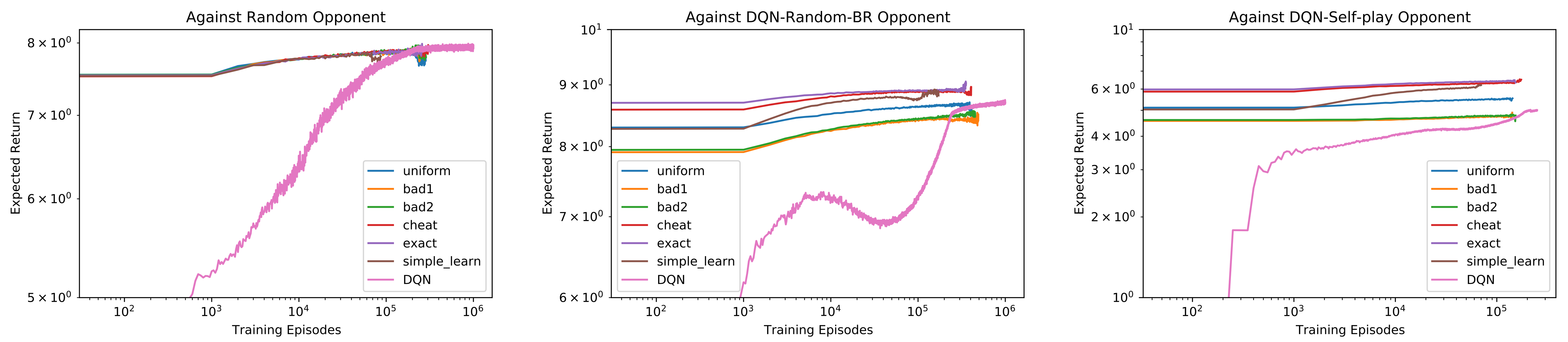
Combining with Population-Based Training

To couple with the non-transitivity and equilibrium selection problem, we combine the new search method with policy-space response oracle (PSRO), which iteratively:

- Compute a distribution over exisiting strategies via empirical game-theoretic analysis
- Compute an approximate best response against this distribution using the search method, and add the new strategy into the pool

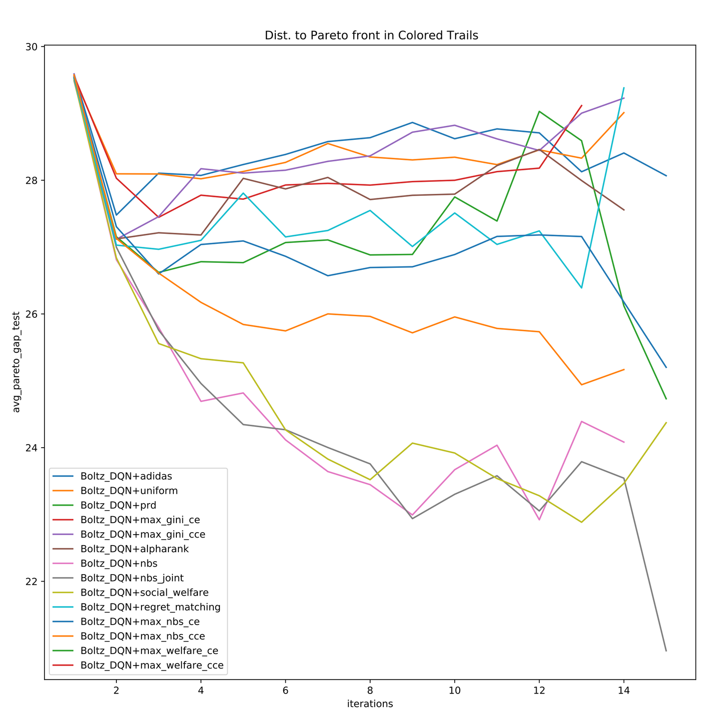


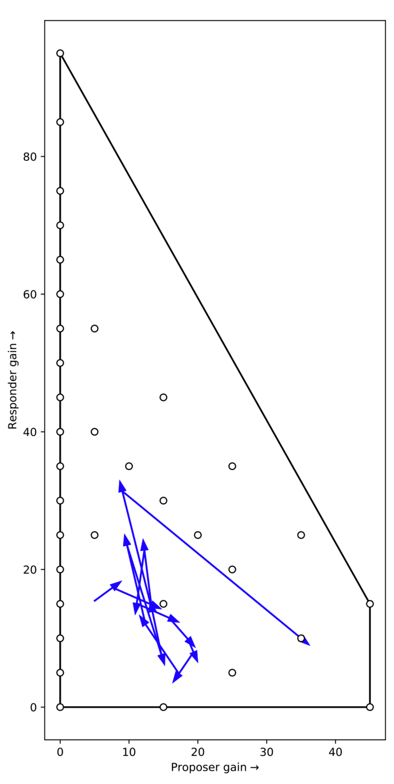
Search-Based Best Response Performances



Nash Bargaining Meta-Strategy Solver

- In the PSRO loop, we can compute the distribution μ as the Nash Bargaining Solution which maximizes players' payoff product: $NBS = \max_\mu \sum_i \log(u_i(\mu) - d_i)$ where $u_i(\mu)$ is the expected payoff under μ, and d_i is the "no-deal" payoff.
- Results on Colored Trails show NBS can reduce the pareto-optimality gap in PSRO loop.





Human-Agent Studies

On DonD game, humans versus agents performance with $N = 129$ human participants, 547 games total. Average performance is given with 95% C.I. in brackets (HvH: 6.93 [6.72, 7.14]).

Agent	$\bar{u}_{\text{Humans}}$	$\bar{u}_{\text{Agent}}$	$\bar{u}_{\text{Comb}}$
IDQN	5.86 [5.37, 6.40]	**6.50** **[5.93, 7.06]**	6.18 [5.82, 6.56]
Comp1	5.14 [4.56, 5.63]	5.49 [4.87, 6.11]	5.30 [4.93, 5.76]
Comp2	6.00 [5.49, 6.55]	5.54 [4.96, 6.10]	5.76 [5.33, 6.12]
Coop	6.71 [6.23, 7.20]	6.17 [5.66, 6.64]	6.44 [6.11, 6.75]
Fair	**7.39** **[6.89, 7.87]**	5.98 [5.44, 6.49]	**6.69** **[6.34, 7.01]**

References

[1] Deal or no deal? End-to-end learning for negotiation dialogues, Lewis, L et. al., EMNLP 2017