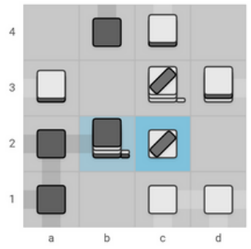


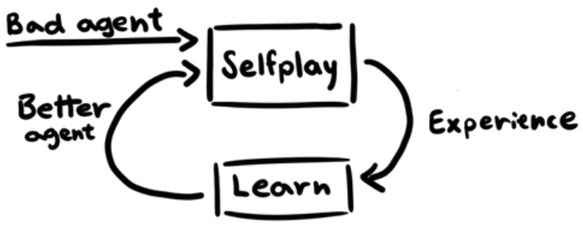
Exploration When Everything Looks New

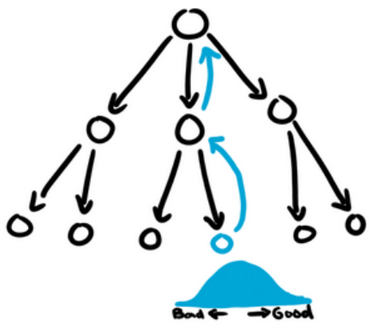
Effect of the Local Uncertainty Source on Exploration

Viliam Vadocz, Matthijs Spaan, Yaniv Oren v.vadocz@student.tudelft.nl TU Delft

Agents improve by interacting with an environment and planning. **By leveraging information about what they don't know, they can learn better and faster**. They do this by estimating the uncertainty in their predictions. There are choices for how to estimate the uncertainty, and we look at what effect this choice has on the exploration and strength of agents playing board games. **We compare the effect of a source of uncertainty which perfectly tracks what the agent has seen (LCGHash), and a source which generalizes (SimHash).**

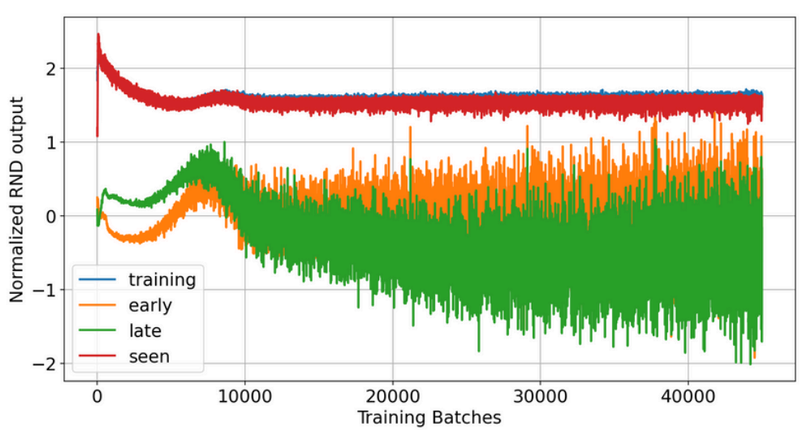
In AlphaZero, agents improve by playing against themselves (selfplay) to generate data about what states are good or bad. They use this data to improve the agent. Importantly, **agents plan** (using MCTS) **during play to act as a better version of themselves.**

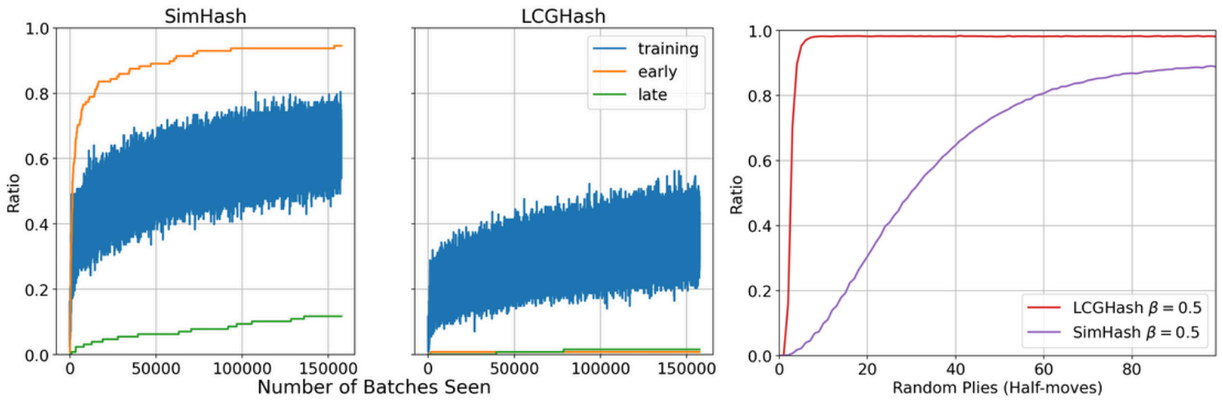




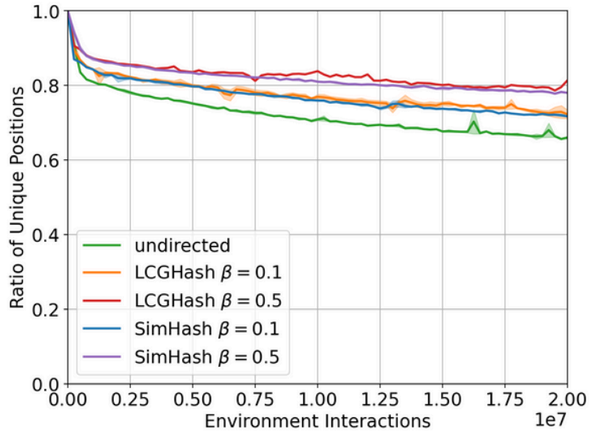
During planning, the agent makes predictions about future states. **These predictions are uncertain when the states have not been seen before.** E-MCTS propagates this uncertainty during planning so that it may be used to explore into new and interesting areas. This uncertainty has to come from some source, which we call the local source of uncertainty.

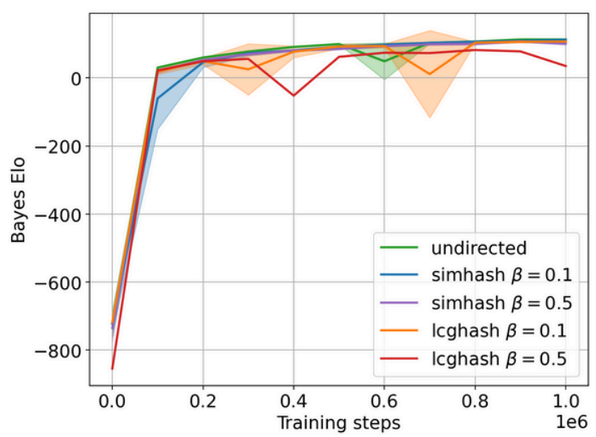
Random Network Distillation (RND) is a popular choice for a local source of uncertainty, but it is very difficult to tune. **A badly tuned RND fails to recognize unseen states**, and so does not provide a reliable source.





Hash-based sources are more reliable, and depending on the hash, we get different behaviour. SimHash generalizes, while LCGHash tracks what states have been seen almost perfectly. **LCGHash is not useful for exploration in board games because the number of states we can reach is too high, so at some point all future states look new and equally interesting.**





Exploring with E-MCTS increases the ratio of unique positions seen throughout training, but does not appear to have an effect on the Elo rating. Higher values of the exploration constant β result in higher ratios. We do not rule out that other sources of uncertainty might produce different behaviour. **Future work should focus on generalizing sources of uncertainty, because truly evidence-based sources like LCGHash fail to provide a useful signal at a certain depth** in very large state spaces, *when everything looks new*.