Information Theory for AI Safety

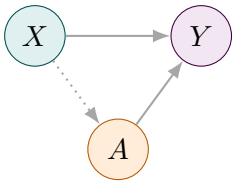
Dovetail Research is offering a paid research fellowship for UK-based researchers interested in uncovering the mathematics behind agents. Here's one problem you could work on.

If we observe a system bringing about a particular outcome, can we conclude that it must be modelling its environment? A theorem by Touchette & Lloyd (2004) shows that if a policy reduces the entropy of its environment beyond what any "blind" policy could achieve, then it must have mutual information with the environment. Consider an environment state X transformed into Y under the influence of an action A (all random variables). Let H denote Shannon entropy and I denote mutual information. Define the entropy reduction achieved by a given policy as $\Delta H = H(X) - H(Y)$, and define ΔH^{Max}_{Blind} as the maximum entropy reduction achievable by any blind policy (where A is independent of X) over any input distribution $\Delta H^{Max}_{Blind} := \max_{P(X),\,P(A)} \Delta H$. Then the Touchette-Lloyd theorem states that:

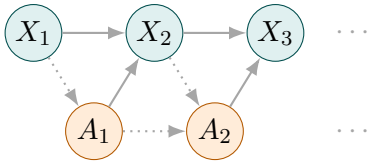
$$\Delta H \leq \Delta H^{Max}_{Blind} + I(X;A)\,.$$

Each additional bit of entropy reduction beyond what the dynamics permit on their own requires at least one bit of mutual information with the environment. However, real agents act over many time-steps. **We wish to extend this theorem to the multi-step setting**, where environment states X_t evolve under actions A_t. In this setting, the action taken by the policy could take information from the previous state or from its previous action (a policy with 'memory'). How much does this extra information help with the task of entropy reduction over multiple timesteps?

Blind vs. sighted (1-step)



Multi-step setup



We wish to investigate this by comparing a policy which is sighted with memory to various other baseline families of policies and bounding their performance in terms of mutual information.

Policy baselines

- **Blind and memoryless:** $P(A_n \mid X_n, A_{n-1}) = P(A_n)$. Each action is drawn i.i.d., independent of both state and action history.
- **Blind with memory:** $P(A_n \mid X_n, A_{n-1}) = P(A_n \mid A_{n-1})$. The current action can depend on its previous action but not on the current environment state.
- **Sighted but memoryless:** $P(A_n \mid X_n, A_{n-1}) = P(A_n \mid X_n)$. The policy can react to the current state but does not use past actions.

What does this have to do with AI Safety?

Understanding how much information is required about the world for an agent to optimize it is important for predicting and controlling AI systems. The Touchette-Lloyd theorem gives us a handle on this problem. Extending it to multiple time-steps would bring it closer to the setting in which real AI systems operate, and could help us reason about how much a system must be modelling its environment given its observed behaviour.

If you have experience in maths, physics, or computer science and are interested in working on problems like this, apply using the QR code. Rolling acceptance until application closes on May 17.

This job is part of an Advanced Research + Invention Agency-funded project.

To see more problems or apply scan here or visit
`dovetailresearch.org`