Fully General Online Imitation Learning

Michael K. Cohen

Marcus Hutter

Neel Nanda

Imitation

- → Optimizers that plan over the long term in an unknown environment are dangerous
- → A perfect imitation of a human is "safe" (not likely to cause an unprecedented catastrophe)
- → Perhaps a population of imitators could prevent the realization of dangerous optimizers
- → Perhaps there could be global coordination to use only imitators and "narrow" optimizers, with too little data to model what makes humans tick

Learned Imitation

- → Errors in the early stages of imitation learning could have irreversible consequences
- → Some concern about dangerousness being *selected for* in the errors of strong prediction algorithms[1]
- → Active learning allows extremely strong error guarantees

[1] Search terms: mesa-optimization; universal prior malign; inner alignment

Online Imitation Learning

- → Instead of train-then-deploy, imitator can ask what the demonstrator would do when uncertain
- → Maintain a posterior distribution over demonstrator-models
- → Sort them
- → Collect the top few. The n^{th} best gets collected if $\texttt{post}(\texttt{model}_n) > \alpha \sum_{i \leq n} \texttt{post}(\texttt{model}_i)$
- → Each model gives probabilities over the next action
- → For each action, take the min prob. assigned by the top models
- → Sample actions with those probs; if nothing sampled, ask for help and take the action the demonstrator picks

Design Motivation

- $\rightarrow$ Suppose the true demonstrator model is always in the top few
- $\rightarrow$ For small enough α, we can make that arbitrarily likely
- $\rightarrow$ a is arbitrary action; θ_q is query prob.
- $\rightarrow$ $\pi^{\mathtt{im}}(a \ \mathtt{w/o\ query} | \mathtt{history}) \leq \pi^{\mathtt{dem}}(a|\mathtt{history})$
- $\rightarrow$ $\pi^{\mathtt{im}}(a \ \mathtt{after\ query} | \mathtt{history}) = \theta_q \pi^{\mathtt{dem}}(a|\mathtt{history})$
- $\rightarrow$ $\pi^{\mathtt{im}}(a|\mathtt{history}) \leq (1+\theta_q)\pi^{\mathtt{dem}}(a|\mathtt{history})$
- $\rightarrow$ actions are never that much more likely for imitator than for demonstrator *even in log space*
- $\rightarrow$ $\mathbb{E}\prod_{t=1}^{N}(1+\theta_q(\mathtt{time\ t})) \in O(N^2)$

(Un)bounded Cost Functions

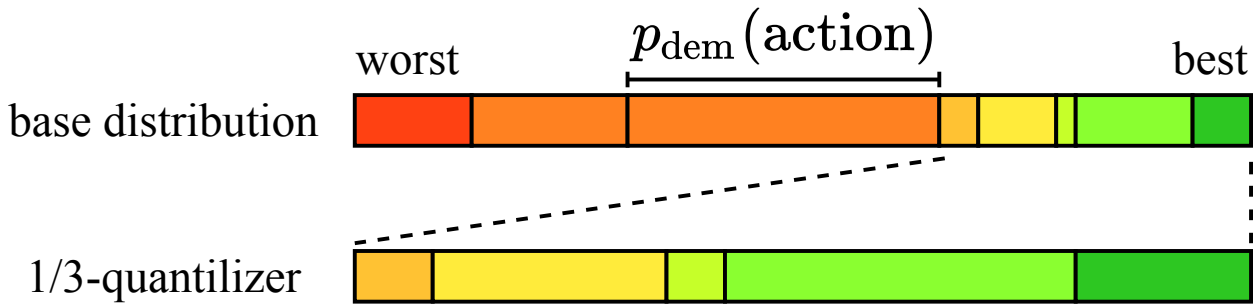
- → Suppose all we had was $||\pi^{\mathtt{im}} - \pi^{\mathtt{dem}}||_1 \leq \varepsilon$
- → What would it take for us to find that satisfactory?
- → Bounded cost function
- → If [bad event] has prob 0 under demonstrator, it has prob $\leq \varepsilon$ under imitator, for added cost $\leq \varepsilon C$
- → Bounded cost function is hard to justify
- → Assume the demonstrator tries to keep bad events unlikely, e.g. bad events $< 10\%$ chance, really bad events $< 1\%$ chance, catastrophic events $< 10^{-6}$
- → Events with 10% chance under demonstrator can only be so bad, maybe
- → 1% events could be much worse
- → An imitator that makes an event go from 1% to 2% way less scary than one that makes it go from 10^{-9} to 1%.
- → Recall: $\pi^{\mathtt{im}}(a|\mathtt{history}) \leq (1+\theta_q)\pi^{\mathtt{dem}}(a|\mathtt{history})$

Main Results

- → Events can unfold over multiple timesteps
- → **Assume:** True demonstrator is in model class
- → Let E be the event that $\texttt{truth}$ always $\in \texttt{top models}$, which has prob. $\geq 1 - \frac{\alpha}{\texttt{prior on truth}}$
- → Let B be a possible bad event in the first t timesteps
- → As $\mathrm{P}_{\texttt{dem}}(B)^{-1} \to \infty$, $\mathrm{P}_{\texttt{im}}(B \text{ and } E)^{-1} \to \infty$ at least polylogarithmicly: $f(x) \gtrsim c \log(x)^3$
- → $c \propto t^{-2}, \alpha^4$
- → $\mathbb{E}\left[\sum_{t=0}^{\infty}(\texttt{query prob})^3\right] \leq |\mathcal{A}|\alpha^{-3}\left(\frac{24}{\texttt{prior on truth}} + 12\right) < \infty$
- → Precondition lifetime with many demonstrator-chosen actions to increase "$\texttt{prior}$" on $\texttt{truth}$

AGI Safety Ideas that Use Imitation

- → **Iterated Distillation and Amplification**
- → Recursive definition: $\texttt{IDA}_0$ is a rock
- → $\texttt{IDA}_{n+1}$ imitates the output of a human with access to two copies of $\texttt{IDA}_n$
- → $\texttt{IDA}_n$ imitates an organization of $2^n - 1$ humans
- → **Quantilization**



- → Quantilization only multiplies action probabilities by a constant factor
- → What if the base distribution were an imitator *without* a bound on $\text{P}_{\text{im}}(\text{action})/\,\text{P}_{\text{dem}}(\text{action})$?
- → The guarantees of quantilization would lose relevance

Conclusion

- → Querying for demonstrations during model-mismatch → finite error bounds in imitation
- → Theoretical solution to safe imitation
- → Theoretical solution to "inner alignment failure"
- → There may be a path to powerful, safe AGI via imitation