# Diverse Hypothesis Set Generation

BY JACK YANSONG LI

The following notes mainly focus on generating a hypothesis set $\mathcal{H}$ that is diverse enough to caputure as many as possible human policies.

## 1 Background

We use the following notations: $\Pi$: set of all policies, $\mathcal{H}^*$: set of all policies adopted by potential partners, and $\mathcal{H}$: hypothesis set that contains the policies generated. The goal of hypothesis set generation is

- *Goal*: Generating $\mathcal{H}$ such that $\mathcal{H}^* \subset \mathcal{H}$.

However, $\mathcal{H}^*$ may contain many redundant policies. We classify these policies by an equivalence relation called *type*. A set of policies that policies belongs to distinct type is call *type-independent* set. The set of all largest type-independent subset of a set $\mathcal{E}$ is denoted as $\mathcal{P}_{\text{type}}(\mathcal{E})$.

- *Goal (given an online glancing algorithm)*: Generating $\mathcal{H}$ such that $\exists \mathcal{H}_{\text{type}} \in \mathcal{P}_{\text{type}}(\mathcal{H}^*)$, $\mathcal{H}_{\text{type}} \subset \mathcal{H}$.

Since we have no information about $\mathcal{H}^*$, it is hard to verify whether the above goal is achieved. However, it is clear if the following goal is achieved, the above goal is also achieved:

- *Opt-Goal (stronger, given an online glancing algorithm)*: Generating $\mathcal{H}$ such that $\exists \mathcal{H}_{\text{type}} \in \mathcal{P}_{\text{type}}(\Pi)$, $\mathcal{H}_{\text{type}} \subset \mathcal{H}$.

However, $|\mathcal{H}_{\text{type}}|$ may be large. Instead, we want to achieve the following suboptimal goal.

- *SubOpt-Goal (weaker, given an online glancing algorithm)*: Generating $\mathcal{H}$ with $|\mathcal{H}| = N$ such that $\mathcal{H}$ is type-independent.

However, it is unclear how "suboptimal" the above goal is given a generated $\mathcal{H}$ with $|\mathcal{H}| = N$. In the following, we give several ideas to formalize the problem and solve it.

## 2 TODO

1. The current definition of type does not match the intuition. A new definition of type should be considered. For example, $\pi \overset{\psi}{\sim} \pi'$ if $\psi$ returns the same best response and $V^*(\pi) = V^*(\pi^*)$.
2. Construct an example of a non online glancing algorithm.
3. It is still unclear what will happen if $\pi^* \notin \mathcal{H}$. The analysis of infinite hypothesis set need to be revised. It will be helpful if the regret bound is a sublinear term adding a linear term that decreases w.r.t. some error. Thus, the error can be use in discretization to generate $\mathcal{H}$. To see this, adopt the result of our current (false) analysis and define $r_{\max} \triangleq \max_{\pi \in \Pi} V(\psi(\pi), \pi)$ and $r_{\min} \triangleq \min_{\pi \in \Pi} V(\psi(\pi), \pi)$. Now, construct a discretization sequence $\{r_j\}_{j=1}^{N+1}$, where $r_1 = r_{\min}$, $r_i < r_j$ if $i < j$, and $r_{N+1} = r_{\max}$. For every $j \in [N+1]$, (Suppose we) can generate a policy $\pi_j$ such that $V(\psi(\pi_j), \pi_j) \in [r_j, r_{j+1}]$[1]. It is clear that $\mathcal{H} \triangleq \{\pi_j\}_{j=1}^{N}$ is type-independent and

$$|V(\psi(\pi_j), \pi_j) - V(\psi(\pi_{j+1}), \pi_{j+1})| \leq \max_{j \in [N+1]} (r_{j+1} - r_j) \quad \forall j \in [N+1].$$

1. In here, the definition of type follows our current definition of type, i.e., $\pi \sim \pi'$ if $V(\psi(\pi), \pi) = V(\psi(\pi'), \pi')$. If we follow the type definition that also requires $\psi(\pi) = \psi(\pi')$ for $\pi \sim \pi'$, we need to add more requirement on the hypothesis generation. For example, instead of generating a single $\pi_j$ such that $V(\psi(\pi_j), \pi_j) \in [r_j, r_{j+1}]$, we need to generate a series of $\{\pi_j^i\}$ such that $V(\psi(\pi_j^i), \pi_j^i) = V(\psi(\pi_j^k), \pi_j^k)$ and $\psi(\pi_j^i) = \psi(\pi_j^k)$ for all $i \neq k$.

Then, the metric $\max_{j\in[N+1]}(r_{j+1}-r_j)$ can be use in an upper bound of the linear term in the regret analysis. The key question of this work is: What the linear term actually depends on? Can we generate a hypothesis set so the factor of the linear term is small?

4. Can we make a switch rule so that the AI agent will switch to a standard RL algorithm such as PPO (starting from $\mu^0=\psi(\pi^*_{\text{fin}})$, where $\pi^*_{\text{fin}}\triangleq\min_{\pi\in\mathcal{H}}|V^*(\pi)-V^*(\pi^*)|$) when she noticed that $\pi^*\notin\mathcal{H}$?